(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
12 December 2002 (12.12.2002)

PCT

(10) International Publication Number
WO 02/099130 A2

(51) International Patent Classification[7]: C12Q 1/68

(21) International Application Number: PCT/GB02/02642

(22) International Filing Date: 7 June 2002 (07.06.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
0113907.0 7 June 2001 (07.06.2001) GB

(71) Applicant *(for all designated States except US)*: **UNIVERSITY COLLEGE LONDON** [GB/GB]; Gower Street, London WC1E 9BT (GB).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **GRIFFITHS, David, John** [GB/GB]; Wohl Virion Centre, Windeyer Institute of Medical Sciences, University College London, 46 Cleveland Street, London W1T 4JF (GB). **KELLAM, Paul** [GB/GB]; Wohl Virion Centre, Windeyer Institute of Medical Sciences, University College London, 46 Cleveland Street, London W1T 4JF (GB). **WEISS, Robert, Anthony** [GB/GB]; Wohl Virion Centre, Windeyer Institute of Medical Sciences, University College London, 46 Cleveland Street, London W1T 4JF (GB).

(74) Agent: **CAMPBELL, Patrick, John, Henry**; J.A. Kemp & Co., 14 South Square, Gray's Inn, London WC1R 5JJ (GB).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 02/099130 A2

(54) Title: VIRUS DETECTION USING DEGENERATE PCR PRIMERS

(57) Abstract: A high throughput method for screening a biological sample for unknown viruses, which method comprises: (a) subjecting DNA from the sample to PCR amplication conditions using simultaneously multiple pairs of degenerate primers, wherein each primer binds a sequence that is conserved across members of a family of viruses and each pair of primers selectively directs amplication of sequence of said family; (b) sequencing PCR product obtained in step (a); and (c) comparing the sequence of the PCR product with the sequences in at least one database comprising viral sequences to determine whether the sequence is present in, or absent from, the database, wherein absence of the sequence from the database suggests that the sequence may be from an unknown virus.

VIRUS DETECTION USING DEGENERATE PCR PRIMERS

Field of the invention

The invention relates to a method of detecting new viruses using a high throughput polymerase chain reaction (PCR) assay.

Background of the invention

Biological materials can often become contaminated or infected with unidentified organisms. For example, cells grown in tissue culture often exhibit signs of a cytopathic effect consistent with a virus infection but the identity of the virus may not be apparent. Human blood products, such as factor VIII for the treatment of haemophiliacs, can be contaminated with unidentified viruses, as was demonstrated by infection of many haemophiliacs with human immunodeficiency virus in the early 1980s. Similarly, two decades ago 20% of individuals who received transfused blood contracted hepatitis C (Randall, 2001, J Pediatr. Oncol. Nurs. 18(1), 4-15).

Summary of the invention

PCR allows amplification of a specific region of a polynucleotide. The specificity of the reaction is due to the primers which, during the course of PCR, bind to the region to be amplified in a sequence specific manner. Degenerate primers can be designed which amplify sequence from substantially all members of a virus family. Such primers typically bind to nucleotide sequence which is conserved across the virus family. The invention provides a PCR based high throughput screen that uses such degenerate primers for detecting unknown viruses.

In particular, the invention provides a high throughput method for screening a biological sample for unknown viruses, which method comprises

(a) subjecting DNA from the sample to PCR amplification conditions using simultaneously multiple pairs of degenerate primers, wherein each primer binds a sequence that is conserved across members of a family of viruses and each pair of primers selectively directs amplification of sequence of said family;

(b) sequencing PCR product obtained in step (a); and

(c) comparing the sequence of the PCR product with the sequences in at least one database comprising viral sequences to determine whether the sequence is present in, or absent from, the database, wherein absence of the sequence from the database suggests that the sequence may be from an unknown virus.

Detailed description of the invention

General description

There are a number of human diseases in which unidentified viruses are thought to play a causative role. For example, unidentified viruses are believed to play a role in cancers such as leukaemia, autoimmune diseases such as rheumatic disease, cardiovascular diseases such as dilated cardiomyopathy and Kawasaki disease, and prostatitis (zur Hausen 2001 The Lancet 357, 381-384; Greaves 1997 The Lancet 349, 344-349; Rowley and Shulman 1998 Clinical Microbiology Reviews 11(3), 405-414; Kawai 1999 Circulation 99, 1091-1100; and Dominigue and Hellstrom 1998 Clinical Microbiology Reviews 11(4), 604-613). Particles resembling retroviruses have been reported in affected tissue from patients with psoriasis, Sjögren's syndrome and rheumatoid arthritis (Iversen 1990 J. Invest. Dermatol. 90, 41S-3S; Garry et al 1990 Science 250, 1127-9; Yamano et al 1997 J. Clin. Pathol. 50, 223-30; and Stransky et al 1993 Br. J. Rheumatol. 32, 1044-8). The invention provides a way of screening for the viruses which may cause or contribute to such diseases. Once identified, the viruses may be used as a target for developing diagnostic tests for, or therapies against, the diseases.

The method of the invention is based on obtaining sequences from viruses so that they can be compared with known viral sequences to determine whether they are from novel viruses. The sequences of the novel viruses are amplified using PCR primers which recognise sequences which are conserved (similar/homologous) in known members of virus families. The primers direct amplification of sequence between the conserved regions to give a PCR product whose sequence can be compared with that of known viruses.

The biological sample which is screened may be any sample susceptible to

infection by a virus. It may, for example, be a tissue culture sample (e.g. tissue culture supernatant), or a sample of animal (including human) or plant material. In a particularly preferred embodiment the invention is directed to the identification of unknown human viruses, and in this case the sample will generally be derived from one or more humans. A sample derived from a human or animal may be from a range of tissue and fluid types, for example blood serum, seminal fluid, breast milk, saliva, cerebrospinal fluid, urine, bile, bronchial lavage fluid, nasal secretion, eye secretion or vaginal wash.

Before the sample is subject to PCR it may processed. In one embodiment the virus material in the sample is concentrated, for example by ultracentrifugation. The virus material may also be purified in a manner which increases the content of viral nucleic acid relative to non-viral nucleic acid. For example the viral nucleic acid may be concentrated by

centrifuging the biological sample under conditions such that cell debris is pelleted and virus particles remain in the supernatant;

collecting the supernatant; and

centrifuging the supernatant under conditions such that virus particles are pelleted.

The initial centrifugation to pellet the cell debris may, for example, be carried out at 100 to 10,000 *g*, preferably from 1000 to 10,000 *g*. The subsequent centrifugation to pellet the virus particles is carried out at a higher *g* force, for example 50,000 to 500,000 *g*, preferably about 100,000 *g*.

The purification of viral nucleic acid may include a step of treating a suspension comprising the virus with a nuclease so as to digest extraneous nucleic acid, wherein the viral nucleic acid is protected from digestion by viral coat or core protein. The nuclease is preferably a non sequence-specific nuclease which digests DNA and/or RNA, for example micrococcal nuclease S7 (Roche Molecular Biochemicals, Catalogue 107 921).

The processing may also comprise a nucleic acid purification, such as phenol/chloroform nucleic acid purification or the use of a column which selectively binds nucleic acid. In one embodiment purification is carried out using a Qiagen™

column.

Typically processing of the sample increases the purity of the virus nucleic acid present in the sample (for example leading to an increase in concentration of 2-fold to 1000-fold of viral nucleic acid).

The processing of the sample may comprise the reverse transcription of viral RNA in the sample to DNA, i.e. RNA from the unknown virus is processed to produce the equivalent (such as the same or a complementary) DNA sequence. Thus the DNA which is subject to PCR conditions may be cDNA. This is required when the unknown virus has an RNA genome. Thus the processing may comprise reverse transcription of the RNA to produce a complementary DNA strand and then optionally synthesising a second DNA strand before carrying out PCR. This can be achieved by using a primer which directs initiation at random sequences in a reverse transcription reaction and then in a second strand synthesis reaction.

Random reverse transcription may be directed using a primer which directs initiation of DNA synthesis at random sequences. Such a primer may be made by synthesising it so that it contains a random sequence, for example a sequence of at least 6 consecutive nucleotides (e.g. from 6 to 20 nucleotides) wherein each nucleotide may be any of the four possible natural nucleotides, i.e. A, T, C or G. In other words, such a primer contains a sequence NNNNNN wherein each N is A, T, C or G.

In one embodiment a "single tube system" is used for the reverse transcription and then PCR with the multiple pairs of degenerate primers. In such a system the sample (typically after being processed) is added to a mixture of reagents which allow both reverse transcription and PCR to occur. Thus typically the mixture will comprise both a reverse transcriptase and a thermostable DNA polymerase. The mixture may comprise the Titan™ reagants from Roche Molecular Biochemicals™ (cat no. 1855476) which uses the avian myeloblastosis virus reverse transcriptase and a Pwo (Pyrococcus woesei) thermostable DNA polymerase. Alternatively the ProSTAR™ system from Stratagene™ may be used.

The PCR reaction is carried out in a PCR mixture that generally comprises the following: the template DNA (which will be amplified in the event of virus

detection), one or more primer pairs specific for members of a virus family, a thermostable polymerase enzyme (typically a DNA polymerase, such as Taq polymerase), deoxynucleotide triphosphates (dATP, dTTP, dCTP and dGTP) and a suitable buffer.

The PCR reaction generally comprises cycles of the following steps: a denaturation step, a primer annealing step and a polynucleotide synthesis step. Typically the PCR reaction comprises at least 25 cycles, such as 30, 35, 40 or more cycles, up to a maximum of 60 cycles for example. Generally, in the denaturation step, the PCR mixture is heated to a temperature at which the DNA in the PCR mixture (in particular the region to be amplified) denatures to single-stranded form. The denaturing temperature is generally from 85 to 98°C.

In one embodiment the PCR reaction comprises a "hot start" in which the PCR mixture is kept at the denaturing temperature for an extended amount of time before commencement of the thermal cycles, such as for 5 to 30 minutes, preferably 10 to 20 minutes. The use of Amplitaq Gold™ DNA polymerase (Applied Biosystems™) is preferred when the PCR reaction comprises a hot start.

In the primer annealing step the primers bind to template nucleotide sequence in a sequence specific manner. This step is generally carried out at a temperature of from 30 to 65°C. In the polynucleotide synthesis step the polymerase replicates/ synthesises nucleotide sequence based on template sequence by addition of nucleotides to the 3' end of the bound primers. This step is generally carried out at about 72°C.

In the method of the invention, the sample (generally after processing as described above) is subject to PCR conditions using a panel of multiple pairs of degenerate primer pairs. In the course of a PCR reaction such primers are capable of binding the conserved sequences of the genome of a family of viruses. These conserved regions typically have a role in providing a necessary or advantageous activity or property to the virus. Generally, the conserved sequences may be coding or non-coding sequences.

In one embodiment the conserved sequences code for or are from virus proteins which have the following activities: DNA or RNA polymerase (replicase),

topoisomerase (helicase/gyrase), endonuclease (integrase), nucleic acid binding protein, protease, transcription factors, envelope glycoproteins, structural protein (e.g. capsid or nucleocapsid protein).

As discussed above multiple pairs of primers are used in the method each of the primer pairs used being selective (or specific) for members of a virus family (for example selective for a subfamily or genus). In the disclosure below regarding the numbers of primers used in different embodiments of the invention it is understood that this refers to the numbers of primers which are substantially specific for members of a virus family. However, in some embodiments additional primer pairs may be used which are selective for more than one family (for example selective for 2 to 10, such as 3 to 6 families). Such embodiments are within the scope of the present invention.

The panel of primer pairs may comprise sets of primer pairs which perform a nested PCR reaction. Generally such a set of primer pairs comprises a first and second primer pair. The first primer pair is able to amplify a template nucleotide sequence from a virus to form a PCR product. The second primer pair is able to amplify a nucleotide sequence using the PCR product generated by the first primer pair as a template. Multiply nested sets of primer pairs may also be used. The use of nested sets of primer pairs allows increased sensitivity and specificity.

The panel of primers used is capable of detecting viruses which are single-stranded or double-stranded DNA or single-stranded or double-stranded RNA viruses. The viruses are generally capable of infecting prokaryotic or eukaryotic cells, such as bacterial, animal, plant, yeast or fungal cells. Preferably the viruses are mammalian (preferably primate) or avian viruses, such as human, pig, horse, sheep, goat, cow, chicken, turkey or duck viruses.

The viruses are typically from any combination of the following families: Adenoviridae, Arenaviridae, Arteriviridae, Astroviridae, Birnaviridae, Bunyaviridae, Caliciviridae, Circoviridae, Coronaviridae, Deltavirus, Filoviridae, Flaviviridae, Hepadnaviridae, Herpesviridae, Orthomyxoviridae, Papovaviridae, Paramyxoviridae, Parvoviridae, Picornaviridae, Polydnaviridae, Poxviridae, Reoviridae, Retroviridae, Rhabdoviridae, Togaviridae or Bornavirus.

Typically in the method 12 to 300 different primer pairs are used, such as 24 to 200 or 48 to 100 primer pairs. These primers may all be used in the same multi-well plate (placed on a thermal cycling machine). The plate may be a 96-well or 384-well plate.

In a preferred embodiment at least one of the wells in which the PCR is done comprises more than one primer pair, such as 2, 3, 4, 5, 6, 7, 8 or 9 primer pairs. Typically 3 to 96, such as 12 to 48, of the wells comprise more than one primer pair.

In one embodiment, some or all of the primer pairs used in the same well carry different labels. Thus, one or both primers of each primer pair carries a label. When both primers of a primer pair carry a label then these labels are different from each other. Typically, at least one of the primers in each primer pair will carry a different label from that used for the other primer pairs in the same well. The PCR product generated by labelled primers carries the labels present on the primers.

Thus, after different primer pairs have been used for PCR in the same well detection of the labels in the PCR products can be used to deduce which primer pair has directed the PCR reaction. In one embodiment all forward primers of the group are labelled with one colour and the reverse primers are labelled with a different colour.

In a preferred embodiment the primers are labelled with a fluorescent label, such as fluorescein based labels (e.g. fluorescein isothiocyanate). Different primer pairs may be labelled with fluorescent labels of different colours. The fluorescent labels which are used may be capable of detection by a Beckman Coulter CEQ2000™ or Applied Biosystems A3700™ fluorescent DNA analyser. The fluorescent labels may be obtained from Beckman Coulter™ or Applied Biosystems™.

Another way of being able to determine which PCR products are generated by which primer pair is for each primer pair in the group to generate a PCR product of different size to the PCR products generated by the other primer pairs of the group. Typically each PCR product which is generated by the group of primers differs in size from all the other PCR products by at least 20, such as at least 50, 100, 200, 500, 1000 or more nucleotides. Each PCR product may for example differ in

size from all other PCR products by up to a maximum of 3000 nucleotides.

In a preferred embodiment, multiple biological samples are screened simultaneously by subjecting DNA from multiple samples to PCR conditions using simultaneously multiple pairs of primers. Generally each of the samples is from a different (typically human) individual. Typically 2 to 80, such as 5 to 40 samples are screened simultaneously in the method.

In one embodiment, DNA from multiple samples is mixed together before being subject to PCR conditions. Typically 2 to 10 such as 5 to 8 samples are pooled together in this way.

After the DNA has been subject to PCR conditions any PCR product which is obtained may be sequenced. Typically prior to sequencing the PCR product is gel purified and cloned into a vector, for example a plasmid or a bacteriophage vector. Suitable plasmids are known and commercially available, such as pBluescript™ (Stratagene) and pGEM-T-Easy™ (Promega). Suitable bacteriophage include bacteriophage λ and M13. Alternatively the sequencing reaction may be carried out on the PCR product itself, for example using one of the PCR primers as a sequencing primer.

Preferably an automated sequencer is used to obtain the sequence of the PCR product, such as a Beckman Coulter CEQ2000™ or Applied Biosystems A3700™ DNA analyser.

Designing the primers

Each of the primer pairs used in the method of the invention binds a sequence conserved across members of a virus family and selectively directs amplification of sequence from the members of the family. The multiple primer pairs which are used are typically designed by:

(i) providing a plurality of amino acid sequences from members of a first virus family,

(ii) comparing the sequences to identify conserved regions,

(iii) designing a first primer pair using a computer based method, wherein each primer in the pair binds a nucleotide sequence that encodes a conserved region

identified in (ii) and wherein the primer pair is designed to amplify by PCR the nucleotide sequence between the nucleotide sequences that encode conserved regions in members of the first virus family, and

(iv) repeating steps (i) to (iii) for each virus family.

The multiple primer pairs may also be designed by:

(i) providing a plurality of nucleotide sequences from members of a first virus family,

(ii) comparing the sequences to identify conserved regions,

(iii) designing a first primer pair using a computer based method, wherein each primer in the pair binds a conserved region identified in (ii) and wherein the primer pair is designed to amplify by PCR the nucleotide sequence between the conserved regions in members of the first virus family, and

(iv) repeating steps (i) to (iii) for each virus family.

The multiple pairs of primers are capable of detecting unknown viruses in a sample, wherein such a sample originates from a single individual or is a pooled sample from individuals of the same species. Thus the panel of primers detects viruses which infect the same species.

The number of primers designed by the above steps is typically the same as the numbers of primers mentioned above for use in the method of the invention. The primer pairs which are designed bind sequence which is conserved across members of a virus family. The panel of primer pairs which is designed may comprise primer pairs that bind sequence which is conserved across substantially members of the family or across a subset of the members of the family, for example across all members of a subfamily or of a genus.

Generally, the primer pairs bind at least 70%, at least 80%, or at least 90% of the known viruses of the family, subfamily or genus. Preferably less than 10, such as less than 5, primer pairs will be used for the detection of any given family, subfamily or genus in the panel.

The panel of primer pairs is generally capable of detecting viruses from at least 10, 15, 20, 30 or more families, typically up to a maximum of 35 families.

The panel of primer pairs may comprise sets of primer pairs which perform a

nested PCR reaction. Generally such a set of primer pairs comprises a first and second primer pair. The first primer pair is able to amplify a template nucleotide sequence from a virus to form a PCR product. The second primer pair is able to amplify a nucleotide sequence using the PCR product generated by the first primer pair as a template. The use of nested sets of primer pairs allows increased sensitivity.

In a preferred embodiment each primer pair is specific for a particular virus family, so that it does not detect viruses of other families.

The plurality of amino acid or nucleotide sequences are provided from different known viruses of the same family. The sequences will be for the same protein of the different viruses. Typically at least 5, 10, 20, 50, 100 or more sequences are provided. The maximum number of sequences provided will, for example, be 300 sequences.

Each of the sequences which is provided is typically at least 20, 50, 100, 200 or more amino acids or nucleotides in length. In general the maximum length of the nucleotide sequences is 1000 nucleotides and the maximum length of the amino acid sequences is 300 amino acids. The sequences may be obtained from a database of sequences, such as GenBank. The sequences may be obtained from a database comprising virus sequences which are organised into homologous protein families (based on sequence similarity relationships).

In a preferred embodiment the sequences are obtained from the VIDA database (described in Alba et al (2001) Nucleic Acids Research 29, 133-136) or the Virus Division of GenBank. The sequences may be provided in the form of a database, preferably in computer-readable form. The sequences are preferably provided in the form of a computer-readable database constructed using programs which identify homologous protein families, such as GeneTableMaker, MKDOM or PSCBuilder.

The sequences which have been provided are compared to identify conserved regions. Typically such conserved regions will have a length of at least 12 nucleotides, such as at least 15, 21, 27, 36, 99 or more nucleotides (generally up to a maximum length of 200 nucleotides) or at least 4, 5, 7, 10, 25 or more amino acids (generally up to a maximum length of 50 amino acids).

Across the conserved region the virus sequences which are being provided will of course share identity or similarity. Typically the amino acids or nucleotides in at least 50% of the positions in the region will be the same in at least 50 %, 60%, 70%, or 80% of the viruses of the group (i.e. in the family, genus or subfamily).

The algorithm which identifies conserved regions generally uses a multiple sequence alignment method. The method may comprise (a) aligning all pairs of sequences separately to calculate a distance matrix giving the divergence of each pair of sequences, (b) calculating a guide tree from the distance matrix, and (c) aligning the sequences progressively according to the branching order in the guide tree.

A preferred algorithm for the aligning the conserved sequences is CLUSTALW as described in Thompson et al (1994) Nucleic Acids Research 22, 4673-80. Other algorithms that can be used for aligning sequences are MultAlin (Corpet (1988) Nucleic Acids Research 16, 10881-90) or Jalview (Clamp et al (1998) http://barton.ebi.co.uk). BLOCKS of conserved regions of amino acids may be extracted from the multiple alignments, typically using the program Blocks Multiple Alignment Processor. Alternatively the entire process of performing multiple alignments and extracting BLOCKS can be performed using BLOCKMAKER (Henikoff and Henikoff (1994) Genomics 19, 97-107).

The output from the alignment and BLOCK extraction set (i.e. the information describing the identified conserved regions) is then entered into the algorithm which designs the primers. Such output is typically in the form of partial sequences which correspond to the conserved regions (BLOCKS). These BLOCKS are input into a primer design algorithm. In one embodiment such an algorithm is CODEHOP.

In the primer design step the conserved regions which are chosen as targets for primers preferably comprise few codons with degenerate counterparts, i.e. preferably the sequence has a low redundancy, such as a redundancy of less than 512 fold, 256 fold or 128 fold. Each primer binds in accordance with Watson-Crick base pairing and thus the binding is sequence specific. Each primer will thus be designed to be wholly or partially complementary to the sequence to which it binds.

Each of the primers typically has a length of at least 8 nucleotides, such as at

least 10, 12, 15, 20, 30, 40 or more nucleotides (up to a maximum of 50 nucleotides for example). In one embodiment the primer may comprises at least 2, 4 or 6, up to a maximum of 10 for example, inosine bases. Inosine is able to bind to any of the four nucleotides and therefore use of inosine causes a reduction in effective redundancy.

Each primer pair will be designed so that the PCR product generally has a length of at least 20, such as at least 50, 100, 200, 500, 1000 or more nucleotides (and typically up to a maximum of $5x10^3$ nucleotides long).

Each primer is preferably be designed so that it anneals to a single site, i.e. the primer will not bind to any other site in the genome of the relevant viruses.

Each primer is preferably designed so that it does not exhibit secondary structure, i.e. the nucleotides in the primer will not bind substantially to any other nucleotide in the primer apart from those to which it is covalently linked. In addition preferably each primer is designed so that it does not bind other primers with the same sequence.

In one embodiment the 3' region, and preferably the 3' terminal nucleotide of the primer binds to the target sequence with high affinity, thus preferably this region or nucleotide comprises a G or C.

Generally each primer is designed to have an annealing temperature of from 30 to 65°C, such as 50 to 60°C or 35 to 45°C. In addition each primer pair may be designed to ensure that the two primers do not bind to each other.

The primers are designed by a computer based algorithm. In one embodiment such an algorithm designs primers according to the following rules:

1) A set of blocks is input, where a block is an aligned array of amino acid sequence segments without gaps that represents a highly conserved region of homologous proteins. A weight is provided for each sequence segment, which can be increased to favour the contribution of selected sequences in designing the primer. A codon usage table is chosen for the target genome.

2) An amino acid position-specific scoring matrix (PSSM) is computed for each block using the odds ratio method.

3) A consensus amino acid residue is selected for each position of the block as the highest scoring amino acid in the matrix.

4) For each position of the block, the most common codon corresponding to the amino acid chosen in step 3 is selected utilizing the user-selected codon usage table. This selection is used for the default 5' consensus clamp in step 8.

5) A DNA PSSM is calculated from the amino acid matrix (step 2) and the codon usage table. The DNA matrix has three positions for each position of the amino acid matrix. The score for each amino acid is divided among its codons in proportion to their relative weights from the codon usage table, and the scores for each of the four different nucleotides are combined in each DNA matrix position. Nucleotide positions are treated independently when the scores are combined. As an option, the highest scoring nucleotide residue from each position can replace the most common codons from step 4 that are used in the consensus clamp.

6) The degeneracy is determined at each position of the DNA matrix based on the number of bases found there. As an option, a weight threshold can be specified such that bases that contribute less than a minimum weight are ignored in determining degeneracy.

7) Possible degenerate core regions are identified by scanning the DNA matrix in the 3' to 5' direction. A core region must start on an invariant 3' nucleotide position, have length of 11 or 12 positions ending on a codon boundary, and have a maximum degeneracy of 128 (this is the default setting of CODEHOP). The degeneracy of a region is the product of the number of possible bases in each position.

8) Candidate degenerate core regions are extended by addition of a 5' consensus clamp from step 4 or 5. The length of the clamp is controlled by a melting point temperature calculation (the CODEHOP default is 60°C) and is usually about 20 nucleotides.

9) Steps 7 and 8 are repeated on the reverse complement of the DNA matrix from step 5 for primers corresponding to the opposite DNA strand.

In one embodiment CODEHOP (Rose et al (1998) Nucleic Acids Research 26, 1628-1635) is used to design the primer pairs. This program uses the above rules.

The primers designed by the algorithm may then be mapped back to the

original sequence to choose primer pairs which provide the desired length of PCR product.

The above-described computer based method is repeated until the desired number of primer pairs have been designed. Optionally the primer pairs can then be synthesised and tested. They are typically tested to determine the optimal conditions for using the primers in a PCR reaction.

In one embodiment the primers are tested for their ability to amplify one or more of the plurality of nucleotide sequences from known viruses which were used to design the primers, or in the case of amino acid sequences from known viruses being used to design the primers the primers may be tested for their ability to amplify the nucleotide sequence from the virus which encodes the amino acid sequence.

The primers may be tested in a range of buffer conditions to determine optimal buffer conditions for PCR using the primers. The buffer conditions which may be tested include pH (typically between 7 and 10), magnesium concentration (typically from 0.5 mM to 5 mM), potassium chloride (typically from 0 to 100 mM), ammonium chloride (typically 0 to 100 mM), glycerol (typically 0 to 20%), dimethysulphoxide (typically 0 to 20%), ethanol (typically 0 to 20%), sorbitol (typically 0 to 20%) or betaine (typically 1M betaine).

The primers may be tested at a range of different temperatures to determine the optimal temperatures in the PCR reaction. Preferably the primers are tested in PCR reaction in which a range of primer annealing temperatures are tested. Typically the range of temperatures is from 30 to 65°C.

The panel of primer pairs or a group of primers within the panel may be designed to be used together on the same plate (i.e. using the same thermal cycles). Thus such primer pairs will be designed to work at the same annealing temperature.

In one embodiment a group of primer pairs within the panel are designed to have similar optimal conditions for use in PCR so that they can be used optimally in the same well or reaction vessel, i.e. that they can be used in multiplex PCR. Such a group typically comprises at least 2, 3, 4, 5, 6 or more primer pairs (up to a maximum of 8 primer pairs for example).

To provide such primer pairs the computer based method steps may be used

to design primer pairs which are calculated to have similar annealing temperatures and/or the primers are tested to select primer pairs which can be used optimally together. Such testing typically determines whether the primers work optimally with the same buffers and/or whether the primers have similar annealing temperatures.

Validating a PCR product as being from a novel virus

After sequencing the PCR product(s), the next step is to determine whether each sequence is present in at least one database of known nucleic acid sequences, typically sequences of viruses known to infect the individuals from which the samples are derived. Appropriate databases include the virus subdivision of GenBank or the VIDA database.

In addition each sequence is typically also compared with a database of human sequences to exclude sequences which are human sequences. Such a database is generally a comprehensive or consensus human genome database. Preferably, at least one of the human sequence databases searched contains an essentially complete human genome sequence. However, it needs to be borne in mind that, although there has recently been a great deal of publicity about the "completion" of the human genome sequence, not all the human genome has in fact been sequenced, and it is possible that a cloned sequence could fall within the unsequenced part of the genome. The human genome contains large areas with repetitive sequences, and much of the unsequenced genome is within these areas.

In order to make as comprehensive a search as possible, it is desirable to search a range of different types of database; in addition to a human genome database, it is desirable to search, for example, a database comprising expressed sequence tags (ESTs) and a database comprising repetitive elements of the human genome. Appropriate databases include GenBank, the EMBL database, the Celera human genome database, the Ensemble human genome database, the DNA Data Bank of Japan (DDBJ), the Incyte LifeSeq™ database of ESTs and the Repbase database of repetitive elements in the human genome.

Where the sequence is found to be not present in any of the interrogated databases of known sequences, this indicates that the nucleic acid may be from a

previously unknown virus. The nucleic acid then becomes a candidate for further investigation and may be designated a Primary Candidate Virus (PCV).

It is generally necessary to confirm by experimentation that a nucleic acid sequence designated a PCV is not in fact a human sequence. A preferred way of doing this involves designing and synthesising a specific primer set (or sets) to amplify the nucleic acid designated a PCV and determining whether the set(s) are able to amplify any DNA in a sample of complete genomic human DNA. The amplification conditions for each primer set may be optimised using the original sample from which the PCV derives or using the PCR product which is obtained in the method of the invention.

The primer set may be used to screen one or more samples of human genomic DNA, for example from 1 to 100 samples, preferably from 5 to 50 samples. As an alternative to PCR, human genomic DNA may be probed with a labelled probe containing sequence from the original PCR product (e.g by Southern blotting).

If the PCV cannot be detected in human DNA by experimentation (by PCR or hybridisation with a labelled probe), it may then be subjected to further analysis. It may be designated a Secondary Candidate Virus (SCV).

The further analysis of an SCV may include gene walking to determine whether the original cloned nucleic acid sequence exists in nature as part of a longer sequence, such as the genomic sequence of an unknown virus. Gene walking may be carried out using techniques known in the art, such as vectorette PCR (Allen et al, PCR Methods Appl. 4:71-75), rapid amplification of cDNA ends (RACE, Frohman et al Proc Natl Acad Sci U S A. 85:8998-9002), rapid amplification of genomic ends (RAGE, Cormack and Somssich. 1997. Gene. 194:273-276) and methods derived from these. Alternatively, the SCV sequence may be "extended" by screening a DNA or cDNA library using the original cloned nucleic acid sequence as a probe.

The additional sequence information obtained through DNA walking may reveal information about the identity of the SCV which cannot be determined from the original clone. The additional information may therefore be analysed, for example to determine whether it contains an open reading frame (i.e. a sequence encoding a protein); the presence of an open reading frame provides further support

for the suggestion that the SCV is a virus. Furthermore, the additional information may identify the SCV as being related to a known virus; for example, the information may identify the SCV as being a new member of a known family of viruses.

A further step may then be to determine whether a newly-identified candidate virus is associated with a disease, for example with a cancer, autoimmune disease, cardiovascular disease or other disease mentioned above. This may be done by obtaining a specimen from each member of a group of subjects with a disease; determining whether the cloned nucleic acid or other nucleic acid of the same virus is present in each specimen; and determining whether the proportion of subjects in whom the nucleic acid is present is greater in the group of subjects who have the disease than in a control group of subjects who do not have the disease, wherein a said greater proportion suggests that the virus may cause or contribute to the disease.

Typically, the process of determining whether the nucleic acid is present or absent from a specimen from a subject may be carried out by PCR using primers specific for the novel sequence (including any contiguous sequence obtained by DNA walking). Initially, perhaps from 10 to 50 patients from a disease group may be tested, but if positive results are obtained in initial studies, the investigation may be extended to a larger group (e.g. a group of up to 100, 500, 1000 or 10,000). The nature of the biological specimens taken from the members of the group varies depending on the disease association that is being investigated; where possible specimens are from disease affected tissue and from peripheral blood of the subjects (for a published example of this see Griffiths et al, 1999, Arthritis Rheumatism, 42:448-454). The specimens may be from the same tissue and fluid types as the biological samples used in the initial screening assay described above.

Once a new virus has been identified and found to be positively associated with a particular disease or condition, serological and genetically-based diagnostic assays for infection by the virus may readily be devised. Genetically-based assays can be developed by using the nucleotide sequence of the virus to design probes and/or PCR primers for specifically detecting the nucleic acid of the virus. Serological assays can be developed by producing recombinant proteins or protein fragments encoded by the virus and testing for the presence of antibodies to these

proteins in human sera. Alternatively, antibodies specific for the proteins of the virus may be made and the antibodies used to detect the virus directly. The serological assays may take the format of an ELISA, western blot or immunofluorescence assay. Correlations may be sought between serological data and genetic data. Furthermore, the organism provides a target for the development of therapies and/or prophylactic vaccines against the disease.

The following Example illustrates the invention:

Example

The Example below refers to Figure 1 which shows how primers were designed using a database known as 'VIDA', and computer programs known as 'CLUSTALW', 'BLOCKS' and 'CODEHOP'.

Designing a panel of primers

A panel of primers was designed for detecting unknown viruses from the family Herpesviridae according to the strategy shown in Figure 1. The amino acid sequences of herpes virus DNA packaging protein UL15 were obtained from the VIDA database (Alba et al, see above). These sequences are shown in Table 1.

The sequences obtained from the VIDA database were then imported into CLUSTALW. This compares the protein sequences to identify conserved regions and then aligns the sequences according to the conserved regions. The alignment produced by CLUSTALW is shown in Table 2.

The BLOCKS program was then used to extract the sequences of the conserved regions identified by CLUSTALW, and to enter these sequences into CODEHOP. The primer sequences were then designed by CODEHOP using the conserved sequences. The output from the CODEHOP program is shown in Table 3.

Table 1. All protein sequences of DNA packaging protein UL15 extracted from VIDA. Here written as a list and unaligned.

>gi_10180719
MFGGLLGEETKRHFERLMKTKNDRLGASHRNERSIRDGDMVDAPFLNFAIPVPRRHQTVMPAIGILHNCC
DSLGIYSAITTRMLYSSIACSEFDELRRDSVPRCYPRITNAQAFLSPMMMRVANSIIFQEYDEMECAAHR
NAYYSTMNSFISMRTSDAFKQLTVFISRFSKLLIASFRDVNKLDDHTVKKRARIDAPSYDKLHGTLELFQ
KMILMHATYFVTSVLLGDHAERAERLLRVAFDTPHFSDIVTRHFRQRATVFLVPRRHGKTWFLVPLIALA
MSSFEGIRIGYTSHIRKAIEPVFEDIGDRLRRWFGAHRVDHVKGETITFSFPSGLKSTVTFASSHNTNSI
RGQDFNLLFVDEANFIRPDAVQTIIGFLNQATCKIIFVSSTNSGKASTSFLYGLKGSADDLLNVVTYICD
EHMKHVTDYTNATSCSCYVLNKPVFITMDGAMRRTAEMFLPDSFMQEIIGGGVVDRTICQGDRSIFTASA
IDRFLIYRPSTVNNQDPFSQDLYVYVDPAFTANTKASGTGVAVIGKYGTDYIVFGLEHYFLRALTGESSD
SIGYCVAQCLIQICAIHRKRFGVIKIAIEGNSNQDSAVAIATRIAIEMISYMKAAVAPTPHNVSFYHSKS
NGTDVEYPYFLLQRQKTTAFDFFIAQFNSGRVLASQDLVSTTVSLTTDPVEYLTKQLTNISEVVTGPTCT
RTFSGKKGGNDDTVVALTMAVYISAHIPDMAFAPIRV
>gi_7673189
MFGGLLGEETKRHFERLMKTKNDRLGASHRNERSIRDGDMVDAPFLNFAIPVPRRHQTVMPAIGILHNCC
DSLGIYSAITTRMLYSSIACSEFDELRRDSVPRCYPRITNAQAFLSPMMMRVANSIIFQEYDEMECAAHR
NAYYSTMNSFISMRTSDAFKQLTVFISRFSKLLIASFRDVNKLDDHTVKKRARIDAPSYDKLHGTLELFQ
KMIFDACHLFCNFCFTWRSRRASERLLRVAFDTPHFSDIVTRHFRQRATVFLVPRRHGKTWFLVPLIALA
MSSFEGIRIGYISHIRKAIEPVFEDIGDRLRRWFGAHRVDHVKGETITFSFPSGLKSTVTFASSHNTNSI
RGQDFNLLFVDEANFIRPDAVQTIIGFLNQATCKIIFVSSTNSGKASTSFLYGLKGSADDLLNVVTYICD
EHMKHVTDYTNATSCSCYVLNKPVFITMDGAMRRTAEMFLPDSFMQEIIGGGVVDRTICQGDRSIFTASA
IDRFLIYRPSTVNNQDPFSQDLYVYVDPAFTANTKASGTGVAVIGKYGTDYIVFGLEHYFLRALTGESSG
SIGYCVAQCLIQICAIHRKRFGVIKIAIEGNSNQDSAVAIATRIAIEMISYMKAAVAPTPHNVSFYHSKS
NGTDVEYPYFLLQRQKTTAFDFFIAQFNSGRVLASQDLVSTTVSLTTDPVEYLTKQLTNISEVVTGPTCT
RTFSGKKGGNDDTVVALTMAVYISAHIPDMAFAPIRV
>gi_5689285
MFGGALGESAKKHFERLLRDRNERLGASRKNECLARGGSLVDAPFLNFAISVPRRHQTVMPAVGTLHDCC
DGTGIYSAIATRLLYAGIVSSEFGEVRRESLSNGHISKRNREALLAPTLTRVANSITFHEYDDAQCAAHR
NAYYSTMNTFGSMRTSDAFQQLASFIDRFSKLLAASFKDVNILDRNNAPKRARITAPSYDKPHGTLELFQ
KMILMHATYFLTSVLLEDHAERAERLLRVIFDIPDFSDAATRHFRQRATVFLVPRRHGKTWFLVPLIALA
MSSFEGIRIGYTSHIRKAIEPVFEEIGDRLRRWFGTQCVDHVKGETITFSFPSGSRSTVTFASSHNTNSI
RGQDFNLLFVDEANFIRPDAVQTIIGFLNQANCKIIFVSSTNSGKASTSFLYGLKGSADDLLNVVTYICD
EHMKHVTNYTNATSCSCYVLNKPVFITMDGAMRRTAEMFLPDSFMKEIIGGITMDRNTCQGDRGVFTASA
VERLLLYRPSTVRNQDILSRDLYVYVDPAFTANTRASGTGIAVIGRYGADYIIFGLEHFFLRALTGESAD
AIGECAAQCIAQICAIHCERFGTIRVAVEGNSNQDSAVAIATRISIDLASYVQSGVAPAPHDVCFYHSKP
AGSNVEYPFFLLQRQKIAAFDFFIARFNSGRVLASQDLVSTTISLSTDPVEYLTKQLTNLSEVVTGATGT
RTFSGKKGGYDDTVVALVMAVYISAHASDATFAPIRGVEATCRGPTEA
>gi_1869837
MFGQQLASDVQQYLERLEKQRQQKVGVDEASAGLTLGGDALRVPFLDFATATPKRHQTVVPGVGTLHDCC
EHSPLFSAVARRLLFNSLVPAQLRGRDFGGDHTAKLEFLAPELVRAVARLRFRECAPEDAVPQRNAYYSV
LNTFQALHRSEAFRQLVHFVRDFAQLLKTSFRASSLAETTGPPKKRAKVDVATHGQTYGTLELFQKMILM
HATYFLAAVLLGDHAEQVNTFLRLVFEIPLFSDTAVRHFRQRATVFLVPRRHGKTWFLVPLIALSLASFR
GIKIGYTAHIRKATEPVFDEIDACLRGWFGSSRVDHVKGETISFSFPDGSRSTIVFASSHNTNGIRGQDF

NLLFVDEANFIRPDAVQTIMGFLNQANCKIIFVSSTNTGKASTSFLYNLRGAADELLNVVTYICDDHMPR
VVTHTNATACSCYILNKPVFITMDGAVRRTADLFLPDSFMQEIIGGQARETGDDRPVLTKSAGERFLLYR
PSTTTNSGLMAPELYVYVDPAFTANTRASGTGIAVVGRYRDDFIIFALEHFFLRALTGSAPADIARCVVH
SLAQVLALHPGAFRSVRVAVEGNSSQDSAVAIATHVHTEMHRILASAGANGPGPELLFYHCEPPGGAVLY
PFFLLNKQKTPAFEYFIKKFNSGGVMASQELVSVTVRLQTDPVEYLSEQLNNLIETVSPNTDVRMYSGKR
NGAADDLMVAVIMAIYLAAPTGIPPAFFPITRTS
>gi_59501
MFGQQLASDVQQYLERLEKQRQLKVGADEASAGLTMGGDALRVPFLDFATATPKRHQTVVPGVGTLHDCC
EHSPLFSAVARRLLFNSLVPAQLKGRDFGGDHTAKLEFLAPELVRAVARLRFKECAPADVVPQRNAYYSV
LNTFQALHRSEAFRQLVHFVRDFAQLLKTSFRASSLTETTGPPKKRAKVDVATHGRTYGTLELFQKMILM
HATYFLAAVLLGDHAEQVNTFLRLVFEIPLFSDAAVRHFRQRATVFLVPRRHGKTWFLVPLIALSLASFR
GIKIGYTAHIRKATEPVFEEIDACLRGWFGSARVDHVKGETISFSFPDGSRSTIVFASSHNTNGIRGQDF
NLLFVDEANFIRPDAVQTIMGFLNQANCKIIFVSSTNTGKASTSFLYNLRGAADELLNVVTYICDDHMPR
VVTHTNATACSCYILNKPVFITMDGAVRRTADLFLADSFMQEIIGGQARETGDDRPVLTKSAGERFLLYR
PSTTTNSGLMAPDLYVYVDPAFTANTRASGTGVAVVGRYRDDYIIFALEHFFLRALTGSAPADIARCVVH
SLTQVLALHPGAFRGVRVAVEGNSSQDSAVAIATHVHTEMHRLLASEGADAGSGPELLFYHCEPPGSAVL
YPFFLLNKQKTPAFEHFIKKFNSGGVMASQEIVSATVRLQTDPVEYLLEQLNNLTETVSPNTDVRTYSGK
RNGASDDLMVAVIMAIYLAAQAGPPHTFAPITRVS
>gi_2605992
MFGKALSRETIQYFETLRKEVQSRSGAKNRAAEAQTGGEDDVKTAFLNFAIPTPQRHQTVVPGVGTLHDC
CETAQIFASVARRLLFRSLSKWRGGESKERLDPSSVEAYVDPKVKQALKTISFVEYNDAEARSCRNAYYS
IMNTFDSLRSSDAFHQVANFVARFSRLVDTSFNGADLDGDGQQTSKRIKVDVPTYGKQRGTLELFQKMIL
MHATYFIAAVILGDHADRIGAFLKMVFNTPEFSDATIRHFRQRATVFLVPRRHGKTWFLVPLIALALATF
KGIKIGYTAHIRKATEPVFDEIGARLRQWFGNSPVDHVKGENISFSFPDGSKSTIVFASSHNTNGIRGQD
FNLLFVDEANFIRPEAVQTIIGFLNQTNCKIIFVSSTNTGKASTSFLYNLKGAADELLNVVTYICDEHME
RVKAHTNATSCSCYILNKPVFITMDGAMRNTAELFLPDSFMQEIIGGGNISGAHRDEPVFTKTAQDRFLL
YRPSTVANQDIMSNNLYVYVDPAFTTNAMASGTGVAVVGRYRSNWIVFGLEHFFLSALTGSSAELIARCV
AQCLAKVFAIHSRPFDSVRIAVEGNSSQDAAVAIATNIQLELNTLRQADVVHMPGTVLFYHCTPPGSSVA
YPFFLLQKQKTGAFDHFIKAFNSGLVLASQELISNTVRLQTDPVEYLLTQMKNLTEVITGTSETRVFTGK
RNGASDDMLVALVMAVYMASLPPTTNAFSSLSTQ
>gi_330792
MFGRVLGRETVQYFEALRREVQARRGAKNRAAEAQNGGEDDAKTAFLNFAIPTPQRHQTVVPGVGTLHDC
CETAQIFASVARRLLFRSLSKWQSGEARERLDPASVEAYVDPKVRQALKTISFVEYSDDEARSCRNAYYS
IMNTFDALRSSDAFHQVASFVARFSRLVDTSFNGADLDGDGQQASKRARVDVPTYGKQRGTLELFQKMIL
MHATYFIAAVILGDHADRIGAFLKMVFNTPEFSDATIRHFRQRATVFLVPRRHGKTWFLVPLIALALATF
KGIKIGYTAHIRKATEPVFDEIGARLRQWFGNSPVDHVKGENISFSFPDGSKSTIVFASSHNTNGIRGQD
FNLLFVDEANFIRPEAVQTIIGFLNQTNCKIIFVSSTNTGKASTSFLYNLKGAADDLLNVVTYICDEHME
RVKAHTNATACSCYILNKPVFITMDGAMRNTAELFLPDSFMQEIIGGGNVSGAHRDEPVFTKTAQDRFLL
YRPSTVANQDIMSSDLYVYVDPAFTTNAMASGTGVAVVGRYRSNWVVFGMEHFFLSALTGSSAELIARCV
AQCLAQVFAIHKRPFDSVRVAVEGNSSQDAAVAIATNIQLELNTLRRADVVPMPGAVLFYHCTPHGSSVA
YPFFLLQKQKTGAFDHFIKAFNSGSVLASQELVSNTVRLQTDPVEYLLTQMKNLTEVVTGTSETRVFTGK
RNGASDDMLVALVMAVYLSSLPPTSDAFSSLPAQ
>gi_971317
MFGGAVGEQSARYFQRLLRERQRRAAERGARPDGGGGARGEDDARVPFLDFAVAAPKRHQTVVPGVGTLH
GYCELAPLFAATASRLLLLTSMARAEAGLNTGTGEAHVSRELAGVLSALRFAAHPPAEAAAHCNAYHSVMA
ALESMRASGAFAQVAAFVARFSRLVGTSFSHLGGGDDADPPRAKRARVEPPSGQTRGALELFQKMILMPA

TYFVAATLLGEHAERIGAFLRVAFNTPDFSDAAVAHFRQRATVFLVPRRHGKTWFLVPLIALALATFKGI
KIGYTAHIRKATEPVFEEIVARLRQWFGGERVDHVKGEVISFSFPDGARSTIVFASSHNTNGIRGQDFNL
LFVDEANFIRPEAVQTIVGFLNQASCKIIFVSSTNTGKASTSFLYNLKGASDGLLNVVTYICNEHTPRVA
AHGGATACSCYVLNKPVFITMDAAARNTAETFLPNSFMQEIIGGGEVARRAEPAAVFTRAAGEQFLLYRP
STAAARGPWPERLYMYIDPAFTSNARASGSGIAVVGRHRGSWLVLGLEHFFLPALTGSSAAEIARCAVRC
FAQVMAVHRRRLDGLFVAVEGNSSQDSAVAIALGVRRELDSLAASGAVPMPAETRFYHCRPPGSAVAYPF
FLLQKQKTAAFDHFIRLFNSGRVVASQDLASLTVRLQTDPVEYLFEQLQNLTESTAGPGGARAFSGKRRG
AADDLMVALVMAVFVGSLPPTDGAFCPLAPRPPAD

>gi_5869808

MSLIMFGRTLGEESVRYFERLKRRRDERFGTLESPTPCSTRQGSLGNATQIPFLNFAIDVTRRHQAVIPG
IGTLHNCCEYIPLFSATARRAMFGAFLSSTGYNCTPNVVLKPWRYSVNANVSPELKKAVSSVQFYEYSPE
EAAPHRNAYSGVMNTFRAFSLSDSFCQLSTFTQRFSYLVETSFESIEECGSHGKRAKVDVPIYGRYKGTL
ELFQKMILMHTTHFISSVLLGDHADRVDCFLRTVFNTPSVSDSVLEHFKQKSTVFLVPRRHGKTWFLVPL
IALVMATFRGIKVGYTAHIRKATEPVFEGIKSRLEQWFGANYVDHVKGESITFSFTDGSYSTAVFASSHN
TNGIRGQDFNLLFVDEANFIRPDAVQTIVGFLNQTNCKIIFVSSTNTGKASTSFLYNLRGSSDQLLNVVT
YVCDDHMPRVLAHSDVTACSCYVLNKPVFITMDGAMRRTADLFMADSFVQEIVGGRKQNSGGVGFDRPLF
TKTARERFILYRPSTVANCAILSSVLYVYVDPAFTSNTRASGTGVAIVGRYKSDWIIFGLEHFFLRALTG
TSSSEIGRCVTQCLGHILALHPNTFTNVHVSIEGNSSQDSAVAISLAIAQQFAVLEKGNVLSSAPVLLFY
HSIPPGCSVAYPFFLLQKQKTPAVDYFVKRFNSGNIIASQELVSLTVKLGVDPVEYLCKQLDNLTEVIKG
GMGNLDTKTYTGKGTTGTMSDDLMVALIMSVYIGSSCIPDSVFMPIK

>gi_5708110

MLGKESVEIVKRYRDALRKRTMERGPDDVDGQEMSDSNFITTASICDRNDSARDTMNSPASRFQFAIDVP
QRHQACIAPIGSFHNCCAISRAFSYMASEIIYENLASYSTKYTDTDAALNDLQVSPKRQLFTGAAEDSIL
PALRQKLANLNFARFAPSDSLIHDKAFDGIMNGYRGFVKSDEFSQLNNFIYRFHTLLKKSFSGQASNDYK
RAKLEKTTSEQRDGTLELFQKMILMHATYFASSICLGEGSTERSNRYLSTVFNTSLFSENIIQHFRQRTT
VFLVPRRHGKTWFLVPLISLLVSSFEGIRIGYTAHLRKATEPVFIEIFTRLYKWFGAKQVEQVKGETITF
TFRNGNKSAIVFASSQNTNGLRGQDFNFLFVDEANFIKPAALHTVMGFLNQTNCKLFFVSSTNTCHSNTS
LLYNLKGKTNSLLNVVTYICDEHMPEIQKRTDVTTCSCYVLHKPVFVSMDSEVRNTADLFVKDSFMHEIA
GGRAGKYDSDRTLVPVRALDQFLIYRPSTSSKPNISGLGKILTVYVDPAFTTNRSASGTGIALVTALRDS
MVLMGAEHFYLDALTGEAALEIAQCVYLCIAYCCLIHAGAFREIRIAVEGNSSQDSAAAIAGNLTELLDS
LRRRLGFSLTFAHSRQPGTAMAHPFYLLNKQKSRAFDLFVSLFNSGRFMASQELVSNTLVLSKDPCEYLV
DQIRNITVTHGQGPDSFRTFSGKQGRVPDDMLVAAVMSTYLALEGSPTAGYHPIAPIGRRQRPA

>gi_1813970

MLRGDSAAKIQERYAELQKRKSHPTSCISTAFTNVATLCRKRYQMMHPELGLAHSCNEAFLPLMAFCGRH
RDYNSPEESQRELLFHERLKSALDKLTFRPCSEEQRASYQKLDALTELYRDPQFQQINNFMTDFKKWLDG
GFSTAVEGDAKAIRLEPFQKNLLIHVIFFIAVTKIPVLANRVLQYLIHAFQIDFLSQTSIDIFKQKATVF
LVPRRHGKTWFIIPIISFLLKHMIGISIGYVAHQKHVSQFVLKEVEFRCRETFARDYVVENKDNVISIDH
RGAKSTALFASCYNTNSIRGQNFHLLLVDEAHFIKKEAFNTILGFLAQNTTKIIFISSTNTTSDSTCFLT
RLNNAPFDMLNVVSYVCEEHLHSFTEKGDATACPCYRLHKPTFISLNSQVRKTANMFMPGAFMDEIIGGT
NKISQNTVLITDQSREEFDILRYSTLNTNAYDYFGKTLYVYLDPAFTTNRKASGTGVAAVGAYRHQFLIY
GLEHFFLRDLSESSEVAIAECAAHMIISVLSLHPYLDELRIAVEGNTNQAAAVRIACLIRQSVQSSTLIR
VLFYHTPDQNHIEQPFYLMGRDKALAVEQFISRFNSGYIKASQELVSYTIKLSHDPIEYLLEQIQNLHRV
TLAEGTTARYSAKRQNRISDDLIIAVIMATYLCDDIHAIRFRVS

>gi_2746296

MLRSCDIDAIQKAYQSIIWKHEQDVKISSTFPNSAIFCQKRFIILTPELGFTHAYCRHVKPLYLFCDRQR
HVKSKIAICDPLNCALSKLKFTAIIEKNTEVQYQKHLELQTSFYRNPMFLQIEKFIQDFQRWICGDFENT

NKKERIKLEPFQKSILIHIIFFISVTKLPTLANHVLDYLKYKFDIEFINESSVNILKQKASVFLVPRRHG
KTWFMIPVICFLLKNLEGISIGYVAHQKHVSHFVMKDVEFKCRRFFPQKNITCQDNVITIEHETIKSTAL
FASCYNTHSIRGQSFNLLIVDESHFIKKDAFSTILGFLPQSSTKIIFISSTNSGNHSTSFLTKLSNSPFE
MLTVVSYVCEDHVHILNDRGNATTCACYRLHKPKFISINADVKKTADLFLEGAFKHEIMGGSLCNVVNDT
LITEQGLIEFDLFRYSTISKQIIPFLGKELYIYIDPAYTINRRASGTGVAAIGTYGDQYIIYGMEHYFLE
SLLSNSDASIAECASHMILAVLELHPFFTELKIIIEGNSNQSSAVKIACILKQTISVIRYKHITFFHTLD
QSQIAQPFYLLGREKRLAVEYFISNFNSGYIKASQELISFTIKITYDPIEYVIEQIKNLHQININEHVTY
NAKKQTCSDDLLISIIMAIYMCHEGKQTSFKEI

>gi_325496

MLRTCDITHIKNNYEAIIWKGERDCSTISTKYPNSAIFYKKRFIMLTPELGFAHSYNQQVKPLYTFCEKQ
RHLKNRKPLTILPSLSHKLQEMKFLPASDKSFESQYTEFLESFKILYREPLFLQIDGFIKDFRKWIKGEF
NDFGDTRKIQLEPFQKNILIHVIFFIAVTKLPALANRVINYLTHVFDIEFVNESTLNTLKQKTNVFLVPR
RHGKTWFIVPIISFLLKNIEGISIGYVAHQKHVSHFVMKEVEFKCRRMFPEKTITCLDNVITIDHQNIKS
TALFASCYNTHQSIRGQSFNLLIVDESHFIKKDAFSTILGFLPQASTKILFISSTNSGNHSTSFLMKLNN
SPFEMLSVVSYVCEDHAHMLNERGNATACSCYRLHKPKFISINAEVKKTANLFLEGAFIHEIMGGATCNV
INDVLITEQGQTEFEFFRYSTINKNLIPFLGKDLYVYLDPAYTGNRRASGTGIAAIGTYLDQYIVYGMEH
YFLESLMTSSDTAIAECAAHMILSILDLHPFFTEVKIIIEGNSNQASAVKIACIIKENITANKSIQVTFF
HTPDQNQIAQPFYLLGKEKKLAVEFFISNFNSGNIKASQELISFTIKITYDPVEYALEQIRNIHQISVNN
YITYSAKKQACSDDLIIAIIMAIYVCSGNSSASFREI

>gi_854039

MKLNNSPFEMLSVVSYVCEDHAHMLNERGNATACSCYRLHKPKFISINAEVKKTANLFLEGAFIHEIMGG
ATCNVINDVLITEQGQTEFEFFRYSTINKNLIPFLGKDLYVYLDPAYTGNRRASGTGIAAIGTYLDQYIV
YGMEHYFLESLMTSSDTAIAECAAHMILSILDLHPFFTEVKIIIEGNSNQASAVKIACIIKENITANKSI
QVTFFHTPDQNQIAQPFYLLGKEKKLAVEFFISNFNSGNIKASQELISFTIKITYDPVEYALEQIRNIHQ
ISVNNYITYSAKKQACSDDLIIAIIMAIYVCSGNSSASFREI

>gi_5733564

MLRTCDITHIKNNYEAIIWKGERNCSTISTKYPNSAIFYKKRFIMLTPELGFAHSYNQQVKPLYTFCEKQ
RHLKNRKPLTILPSLTRKLQEMKFLPASDKSFESQYTEFLESFKILYREPLFLQIDGFIKDFRKWIKGEF
NDFGDTRKIQLEPFQKNILIHVIFFIAVTKLPALANRVINYLTHVFDIEFVNESTLNTLKQKTNVFLVPR
RHGKTWFIVPIISFLLKNIEGISIGYVAHQKHVSHFVMKEVEFKCRRMFPEKTITCLDNVITIDHQNIKS
TALFASCYNTHSIRGQSFNLLIVDESHFIKKDAFSTILGFLPQASTKILFISSTNSGNHSTSFLMKLNNS
PFEMLSVVSYVCEDHAHMLNERGNATACSCYRLHKPKFISINAEVKKTANLFLEGAFIHEIMGGATCNVI
NDVLITEQGQTEFEFFRYSTINKNLIPFLGKDLYVYLDPAYTGNRRASGTGIAAIGTYLDQYIVYGMEHY
FLESLMTSSDTAIAECAAHMILSILDLHPFFTEVKIIIEGNSNQASAVKIACIIKENITANKSIQVTFFH
TPDQNQIAQPFYLLGKEKKLAVEFFISNFNSGNIKASQELISFTIKITYDPVEYALEQIRNIHQISVNNY
ITYSAKKQACSDDLIIAIIMAIYVCSGNSSASFREI

>gi_4996048

MKLNNSPFEMLSVVSYVCEDHAHMLNERGNATACSCYRLHKPKFISINAEVKKTANLFLEGAFIHEIMGG
ATCNVINDVLITEQGQTEFEFFRYSTINKNLIPFLGKDLYVYLDPAYTGNRRASGTGIAAIGTYLDQYIV
YGMEHYFLESLMTSSDTAIAECAAHMILSILDLHPFFTEVKIIIEGNSNQASAVKIACIIKENITANKSI
QVTFFHTPDQNQIAQPFYLLGKEKKLAVEFFISNFNSGNIKASQELISFTIKITYDPVEYALEQIRNIHQ
ISVNNYITYSAKKQACSDDLIIAIIMAIYVCSGNSSASFREI

>gi_1136808

MLLSRHRERLAANLEETAKDAGERWELSAPTFTRHCPKTARMAHPFIGVVHRINSYSSVLETYCTRHHPA
TPTSANPDVGTPRPSEDNVPAKPRLLESLSTYLQMRCVREDAHVSTADQLVEYQAGRKTHDSLHACSVYR
ELQAFLVNLSSFLNGCYVPGVHWLEPFQQQLVMHTFFFLVSIKAPQKTHQLFGLFKQYFGLFETPNSVLQ

TFKQKASVFLIPRRHGKTWIVVAIISMLLASVENINIGYVAHQKHVANSVFAEIIKTLCRWFPPKNLNIK
KENGTIIYTRPGGRSSSLMCATCFNKNSIRGQTFNLLYVDEANFIKKDALPAILGFMLQKDAKLIFISSV
NSSDRSTSFLLNLRNAQEKMLNVVSYVCADHREDFHLQDALVSCPCYRLHIPTYITIDESIKTTTNLFME
GAFDTELMGEGAASSNATLYRVVGDAALTQFDMCRVDTTAQEVQKCLGKQLFVYIDPAYTNNTEASGTGV
GAVVTSTQTPTRSLILGMEHFFLRDLTGAAAYEIASCACTMIKAIAVLHTTIERVNAAVEGNSSQDSGVA
IATVLNEICPLPIHFLHYTDKSSALQWPIYMLGGEKSSAFETFIYALNSGTLSASQTVVSNTIKISFDPV
TYLVEQVRAIKCVPLRDGGQSYSAKQKHMSDDLLVAVVMAHFMATDDRHMYKPISPQ

>gi_1718281

MLQKDAKLIFISSVNSSDRSTSFLLNLRNAQEKMLNVVSYVCADHREDFHLQDALVSCPCYRLHIPTYIT
IDESIKTTTNLFMEGAFDTELMGEGAASSNATLYRVVGDAALTQFDMCRVDTTAQQVQKCLGKQLFVYID
PAYTNNTEASGTGVGAVVTSTQTPTRSLILGMEHFFLRDLTGAAAYEIASCACTMIKAIAVLHPTIERVN
AAVEGNSSQDSGVAIATVLNEICPLPIHFLHYTDKSSALQWPIYMLGGEKSSAFETFIYALNSGTLSASQ
TVVSNTIKISFDPVTYLVEQVRAIKCVPLRDGGQSYSAKQKHMSDDLLVAVVMAHFMATDDRHMYKPISP
Q

>gi_2246515

MLQKDAKLIFISSVNSSDRSTSFLLNLRNAQEKMLNVVSYVCADHREDFHLQDALVSCPCYRLHIPTYIT
IDESIKTTTNLFMEGAFDTELMGEGAASSNATLYRVVGDAALTQFDMCRVDTTAQQVQKCLGKQLFVYID
PAYTNNTEASGTGVGAVVTSTQTPTRSLILGMEHFFLRDLTGAAAYEIASCACTMIKAIAVLHPTIERVN
AAVEGNSSQDSGVAIATVLNEICPLPIHFLHYTDKSSALQWPIYMLGGEKSSAFETFIYALNSGTLSASQ
TVVSNTIKISFDPVTYLVEQVRAIKCVPLRDGGQSYSAKQKHMSDDLLVAVVMAHFMATDDRHMYKPISP
Q

>gi_2246552

MLLSRHRERLAANLQETAKDAGERWELSAPTFTRHCPKTARMAHPFIGVVHRINSYSSVLETYCTRHHPA
TPTSANPDVGTPRPSEDNVPAKPRLLESLSTYLQMRCVREDAHVSTADQLVEYQAARKTHDSLHACSVYR
ELQAFLVNLSSFLNGCYVPGVHWLEPFQQQLVMHTFFFLVSIKAPQKTHQLFGLFKQYFGLFETPNSVLQ
TFKQKASVFLIPRRHGKTWIVVAIISMLLASVENINIGYVAHQKHVANSVFAEIIKTLCRWFPPKNLNIK
KENGTIIYTRPGGRSSSLMCATCFNKNSIRGQTFNLLYVDEANFIKKDALPAILGFMLQKDAKLIFISSV
NSSDRSTSFLLNLRNAQEKMLNVVSYVCADHREDFHLQDALVSCPCYRLHIPTYITIDESIKTTTNLFME
GAFDTELMGEGAASSNATLYRVVGDAALTQFDMCRVDTTAQQVQKCLGKQLFVYIDPAYTNNTEASGTGV
GAVVTSTQTPTRSLILGMEHFFLRDLTGAAAYEIASCACTMIKAIAVLHPTIERVNAAVEGNSSQDSGVA
IATVLNEICPLPIHFLHYTDKSSALQWPIYMLGGEKSSAFETFIYALNSGTLSASQTVVSNTIKISFDPV
TYLVEQVRAIKCVPLRDGGQSYSAKQKHMSDDLLVAVVMAHFMATDDRHMYKPISPQ

>gi_4494933

MLQKDAKLIFISSSNSSDKSTSFLLNLKDAHEKMLNVVNYVCPDHKDDFNLQDTVVACPCYRLHIPAYIT
IDETVRSTTNLFLEGAFSTELMGDAATSAQSMHKIVSDSSLSQLDLCRVKSTSQDIQGAMKPCLHVYIDP
AYTNNTDASGTGIGAVIAVNHKVIKCILLGVEHFFLRDLTGTAAYQIASCAAALIRAIVTLHPQITHVNV
AVEGNSSQDAGVAIATVLNEICSVPLSFLHHVDKNTLIRSPIYMLGPEKAKAFESFIYALNSGTFSASQT
VVSHTIKLSFDPVAYLIDQIKAIRCIPLKDGGHTYCAKQKTMSDDVLVAAVMAHYMATNDKFVFKSLE

>gi_7330018

MLQKDAKLIFISSSNSSDKSTSFLLNLKDAHEKMLNVVNYVCPDHKDDFNLQDTVVACPCYRLHIPAYIT
IDETVRSTTNLFLEGAFSTELMGDAATSAQSMHKIVSDSSLSQLDLCRVESTSQDIQGAMKPCLHVYIDP
AYTNNTDASGTGIGAVIAVNHKVIKCILLGVEHFFLRDLTGTAAYQIASCAAALIRAIVTLHPQITHVNV
AVEGNSSQDAGVAIATVLNEICSVPLSFLHHADKNTLIRSPIYMLGPEKAKAFESFIYALNSGTFSASQT
VVSHTIKLSFDPVAYLIDQIKAIRCIPLKDGGHTYCAKQKTMSDDVLVAAVMAHYMATNDKFVFKSLE

>gi_4019255

MLLLKAKKALMENLTEASSTQSETEWTVDTPTMITNIKKSERMAYSKIGVIPSINLYSASLTSFCRLYRP

LALKQPLPQTGTLRLLPSEKPYISQKLSNYVKSLTLKHVWHDIEAEAEYYASVQTEKTFMECPIYLELRQ
FIINLSSFLNGCYVKKSTHIEPFQLQLILHTFYFLISIKSPESTNKLFDIFKEYFGLGEMDSAMLQNFKQ
KASIFLIPRRHGKTWIVVAIISMLITSVENLHVGYVAHQKHVANSVFTEIINTLQKWFPSKNIDVKKENG
TIIYKIPGKKPSTLMCASCFNKNSIRGQTFNLLYIDEANFIKKDSLPAILGFMLQKDAKLIFISSVNSGD
KATSFLFNLKNASEKMLNIVNYICPDHKDDFSLQDSLISCPCYKLYIPTYITIDETIKNTTNLFLDGAFT
TELMGDISVMSKNNIHKVIGETALMQFDLCRIDTTKPEITQCLNSIMYLYIDPAYTNNSEASGTGIGAII
ALKNNSSKCIIVGIEHYFLKDLTGTATYQIASCACSLIRAALVLYPHIQAVHVAVEGNSSQDSAVAISTF
LNECSPVKVNFMHYKDKTTAMQWPIYMLGSEKSQAFESFIYAINSGTISASQSIISNTIKLTFDPISYLI
EQIRAIRCYPLRDGSHTYCAKKRTVSDDVLVAVVMAHFFSTSNKHIFKQLNSI
>gi_4019257
MLQKDAKLIFISSVNSGDKATSFLFNLKNASEKMLNIVNYICPDHKDDFSLQDSLISCPCYKLYIPTYIT
IDETIKNTTNLFLDGAFTTELMGDISVMSKNNIHKVIGETALMQFDLCRIDTTKPEITQCLNSIMYLYID
PAYTNNSEASGTGIGAIIALKNNSSKCIIVGIEHYFLKDLTGTATYQIASCACSLIRAALVLYPHIQAVH
VAVEGNSSQDSAVAISTFLNECSPVKVNFMHYKDKTTAMQWPIYMLGSEKSQAFESFIYAINSGTISASQ
SIISNTIKLTFDPISYLIEQIRAIRCYPLRDGSHTYCAKKRTVSDDVLVAVVMAHFFSTSNKHIFKQLNS
I
>gi_60355
MLLLKAKKAIIENLSEVSSTQAETDWDMSTPTIITNTSKSERTAYSKIGVIPSVNLYSSTLTSFCKLYHP
LTLNQTQPQTGTLRLLPHEKPLILQDLSNYVKLLTSQNVCHDTEANTEYNAAVQTQKTSMECPTYLELRQ
FVINLSSFLNGCYVKRSTHIEPFQLQLILHTFYFLISIKSPESTNRLFDIFKEYFGLREMDPDMLQIFKQ
KASIFLIPRRHGKTWIVVAIISMLLTSVENIHVGYVAHQKHVANSVFTEIINTLQKWFPSRYIDIKKENG
TIIYKSPDKKPSTLMCATCFNKNSIRGQTFNLLYIDEANFIKKDSLPAILGFMLQKDAKLIFISSVNSGD
RATSFLFNLKNASEKMLNIVNYICPDHKDDFSLQDSLISCPCYKLYIPTYITIDETIKNTTNLFLDGAFT
TELMGDMSGISKSNMHKVISEMAITQFDLCRADTTKPEITQCLNSTMYIYIDPAYTNNSEASGTGIGAIL
TFKNNSSKCIIVGMEHYFLKDLTGTATYQIASCACSLIRASLVLYPHIQCVHVAVEGNSSQDSAVAISTL
INECSPIKVYFIHYKDKTTTMQWPIYMLGAEKSIAFESFIYAINSGTISASQSIISNTIKLSFDPISYLI
EQIRSIRCYPLRDGSHTYCAKKRTVSDDVLVAVVMAYFFATSNKHIFKPLNST
>gi_695201
MLQKDAKIIFISSVNSSDQTTSFLYNLKNAKEKMLNVVNYVCPQHREDFSLQESVVSCPCYRLHIPTYIA
IDENIKDTTNLFMEGAFTTELMGDGAAATTQTNMHKVVGEPALVQFDLCRVDTGSPEAQRGLNPTLFLYV
DPAYTNNTEASGTGMGAVVSMKNSDRCVVVGVEHFFLKELTGASSLQIASCAAALIRSLATLHPFVREAH
VAIEGNSSQDSAVAIATLLHERSPLPVKFLHHADKATGVQWPMYILGAEKARAFETFIYALNSNTLSCGQ
AIVSNTIKLSFDPVAYLIEQIRAIKCYPLKDGTVSYCAKHKGGSDDTLVAVVMAHYFATSDRHVFKNHMK
QI
>gi_4928934
MLLSSFRNHLQKNYEKYSVQAQNIDWPVETPVLISKDSKTNRLAHPLIGVISRINLYSPTLKYYCDEYST
TKQPKFTPDIGYVRDLKKHDQYFLPKLQHHLSTLCEAYNHVDRQAQVEFNASILTLKAFNANGVLNELKQ
FLINLSCFLNGCYVSKSTCIELFQKQLILHTFYFLISIKTPEETNKMFTFFKHYVGLFDIDDNMLQCFKQ
KSTVFLIPRRHGKTWIVVAIISVLLASVENVHIGYVAHQKHVANAVFTEIITTLYQWFPSKNIEIKKENG
TIIYTKPGRKPSTLMCATCFNKNSIRGQTFNILYVDEANFIKKEALPAILGFMLQKDAKIIFISSVNSAD
KSTSFLFNLRNAKEKMLNVVNYVCPEHKEDFNLQSTLTSCPCYRLHIPTYITIDESIKNTTNLFLDDVFT
TELMGDISTFPTSSMFKVVEEQALFHFDICRVDTTQIDTVKIIDNVLYVYVDPAYTSNSEASGTGIGAVV
PLKTKVKTIILGIEHFYLKNLTGTASQQIAYCVTSMIKAILTLHPHINHVNVAVEGNSSQDSAVAISTFI
NEYCPVPVFFAHCNERSSVFQWPIYILGSEKSQAFEKFICALNTGTLSASQTIVSNTIKISFDPVAYLME
QIRAIRCLPLKDGSYTYCAKQKTMSDDTLVAVVMANYMAISEKHTFKELCKT
>gi_1632798

MLYASQRGRLTENLRNALQQDSTTQGCLGAETPSIMYTGAKSDRWAHPLVGTIHASNLYCPMLRAYCRHY
GPRPVFVASDESLPMFGASPALHTPVQVQMCLLPELRDTLQRLLPPPNLEDSEALTEFKTSVSSARAILE
DPNFLEMREFVTSLASFLSGQYKHKPARLEAFQKQVVLHSFYFLISIKSLEITDTMFDIFQSAFGLEEMT
LEKLHIFKQKASVFLIPRRHGKTWIVVAIISLILSNLSNVQIGYVAHQKHVASAVFTEIIDTLTKSFDSK
RVEVNKETSTITFRHSGKISSTVMCATCFNKNSIRGQTFHLLFVDEANFIKKEALPAILGFMLQKDAKII
FISSVNSADQATSFLYKLKDAQERLLNVVSYVCQEHRQDFDMQDSMVSCPCFRLHIPSYITMDSNIRATT
NLFLDGAFSTELMGDTSSLSQGSLSRTVRDDAINQLELCRVDTLNPRVAGRLASSLYVYVDPAYTNNTSA
SGTGIAAVTHDRADPNRVIVLGLEHFFLKDLTGDAALQIATCVVALVSSIVTLHPHLEEVKVAVEGNSSQ
DSAVAIASIIGESCPLPCAFVHTKDKTSSLQWPMYLLTNEKSKAFERLIYAVNTASLSASQVTVSNTIQL
SFDPVLYLISQIRAIKPIPLRDGTYTYTGKQRNLSDDVLVALVMAHFLATTQKHTFKKVH
>gi_2337991
MFYVKVMPALQKACEELQNQWSAKSGKWPVPETPLVAVETRRSERWPHPYLGLLPGVAAYSSTLEDYCHL
YNPYIDALTRCDLGQTHRRVATQPVLSDQLCQQLKKLFSCPRNTSVKAKLEFEAAVRTHQALDNSQVFLE
LKTFVLNLSAFLNKRYSDRSSHIELFQKQLIMHTFFFLVSIKAPELCEKFCNIFKLYFNIDTMDQATLDI
FKQKASVFLIPRRHGKTWIVVAIISILLASVQDLRIGYVAHQKHVANAVFTEVINTLHTFFPGKYMDVKK
ENGTIIFGLPNKKPSTLLCATCFNKNSIRGQTFQLLFVDEANFIKKDALPTILGFMLQKDAKIIFISSSN
SSDQSTSFLYNLKGASERMLNVVSYVCSNHKEDFSMQDGLISCPCYSLHVPSYISIDEQIKTTTNLFLDG
VFDTELMGDSSCGTLSTFQIISESALSQFELCRIDTASPQVQAHLNSTVHMYIDPAFTNNLDASGTGISV
IGRLGAKTKVILGCEHFFLQKLTGTAALQIASCATSLLRSVVIIHPMIKCAQITIEGNSSQDSAVAIANF
IDECAPIPVTFYHQSDKTKGVLCPLYLLGQEKAVAFESFIYAMNLGLCKASQLIVSHTIKLSFDPVTYLL
EQVRAIKCQSLRDGSHTYHAKQKNLSDDLLVSVVMSLYLSSANTLPFKPLHIERFF
>gi_2317977
MLQKDAKIFFISSVNSGEKTTSFLYNLKDANEKMVNVVSYVCSEHMEDFNKQSAITACPCYRLYVPEFIT
INDNIKCTTNLLLEGSFATELMGNMQSHTEVSGNSMIHESSLTRLDFYRCDTAGQGAPTTENTLFVYIDP
AYGNNVHASGTGIVAMSHCKHTKKCIILGLEHFFLNNLTGTAAHNIASCATALLEGILFQHPWIQEIRCI
IEGNSNQDSAVAIATFISHNIKLPTLFASYRDKTGMQWPIYMLSGDKTLAFQNFISSLNQGLLCASQTVV
SNTVLLSSDPISYLIEQIKNTKCIYHKNKTITFQSKTHTMSDDVLIACVMTCYVMTTNKISYISFSIK
>gi_6625593
MFIASKKSYFEAVYRSTVSSHSEEFWKSDDPVYFTQYKKQCNRLPNAYLGTLHSASKYSENFRHYVATFS
NSPLDFPQSVFNERNPCEYSVPYLDSALQCSAKTLVGCSVSTTERNEYEVCKEATRCFKDAMSHKVLKVF
LSNLSWFLKGHYKSKQAFLEPFQKQLILHSFMFVASIKCPETTTKLFDEFKFLLDMLYFDNTDLLTFLQK
SPAFLIPRRHGKTWIVTAIISMLLTSVDDLHIGYVAHQKHVSLAVFLEISNILLAWFPRKNIDIKKENGV
ILYSHPGKKSSTLMCATCFNKNSIRGQTFNLLFVDEANFIKKEALPAILGFMLQKDAKIFFISSVNSGEK
TTSFLYNLKDANEKMVNVVSYVCSEHMEDFNKQSAITACPCYRLYVPEFITINDNIKCTTNLLLEGSFAT
ELMGNMQSHTEVSGNSMIHESSLTRLDFYRCDTAGQGAPTTENTLFVYIDPAYGNNVHASGTGIVAMSHC
KHTKKCIILGLEHFFLNNLTGTAAHNIASCATALLEGILFQHPWIQEIRCIIEGNSNQDSAVAIATFISH
NIKLPTLFASYRDKTGMQWPIYMLSGDKTLAFQNFISSLNQGLLCASQTVVSNTVLLSSDPISYLIEQIK
NTKCIYHKNKTITFQSKTHTMSDDVLIACVMTCYVMTTNKISYISFSIK

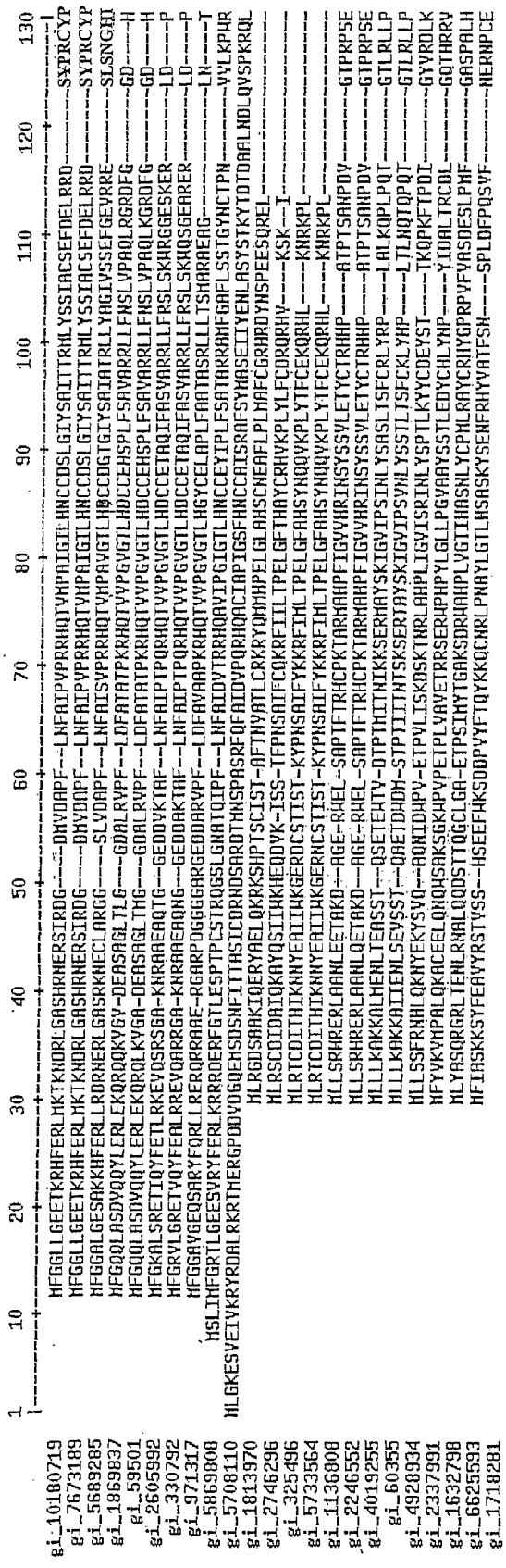


TABLE 2

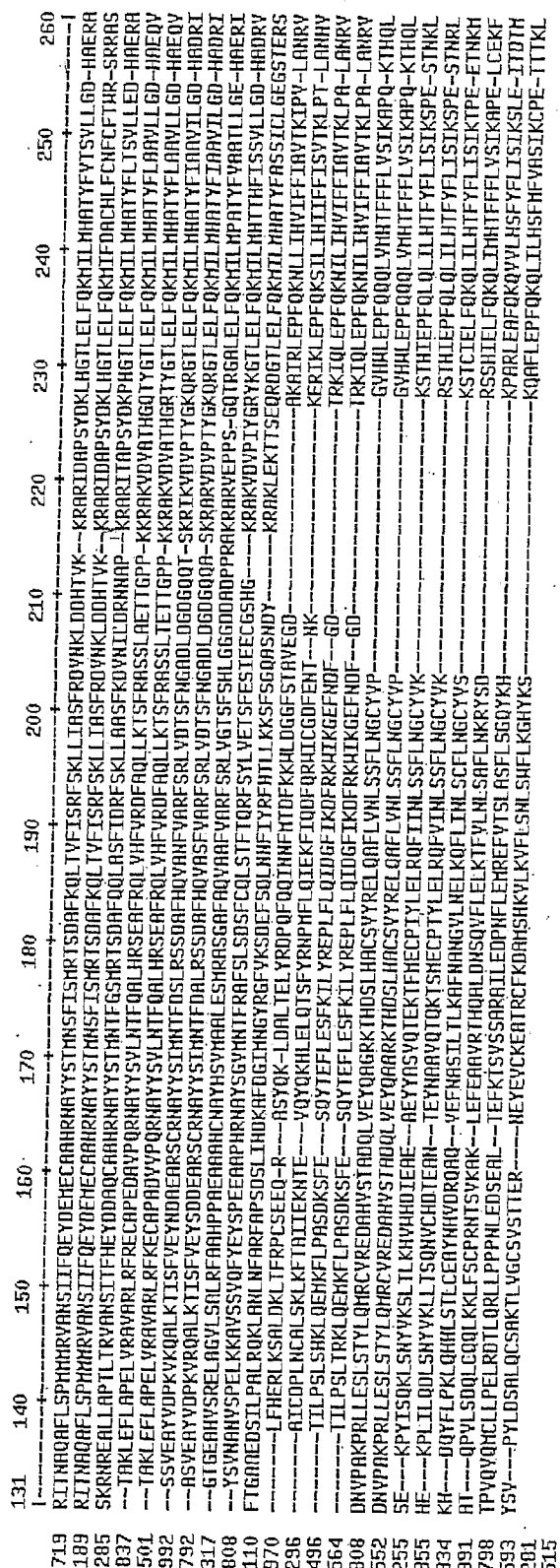


TABLE 2 CONTINUED

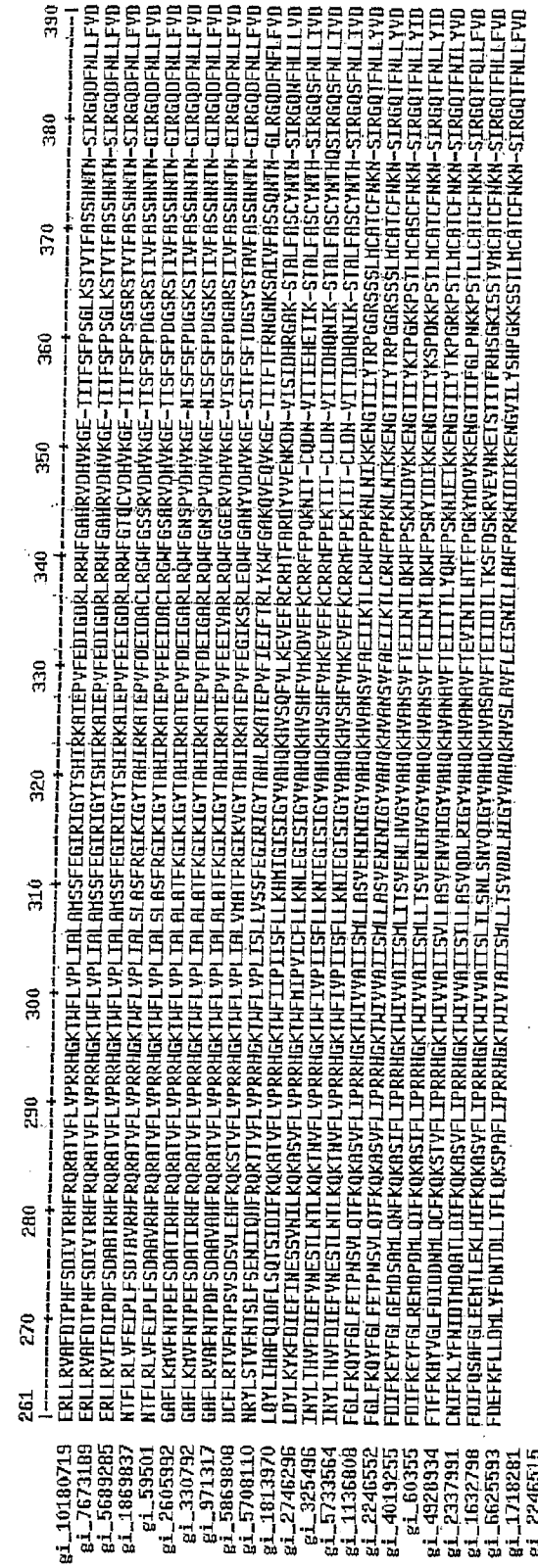


TABLE 2 CONTINUED

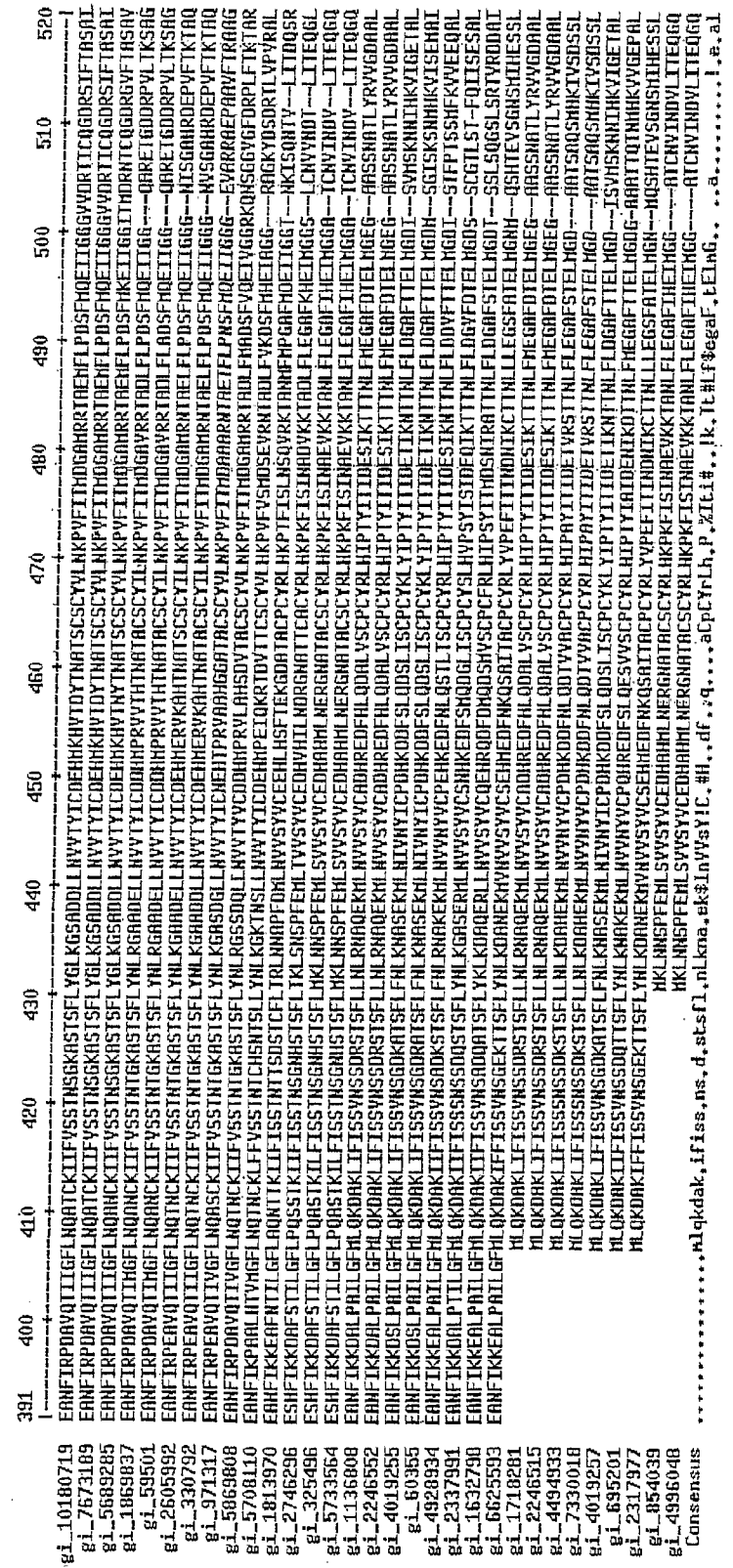


TABLE 2 CONTINUED

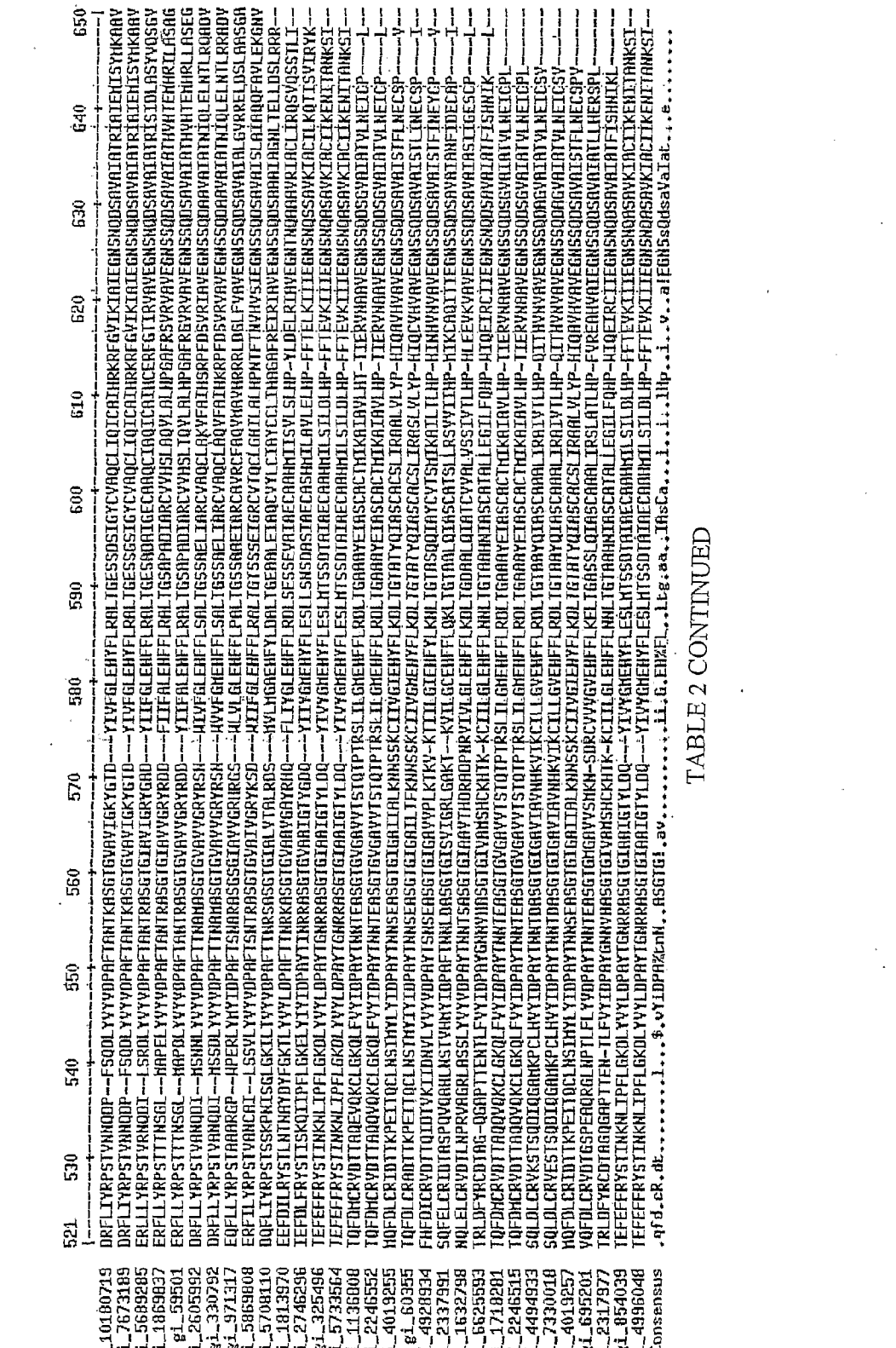


TABLE 2 CONTINUED

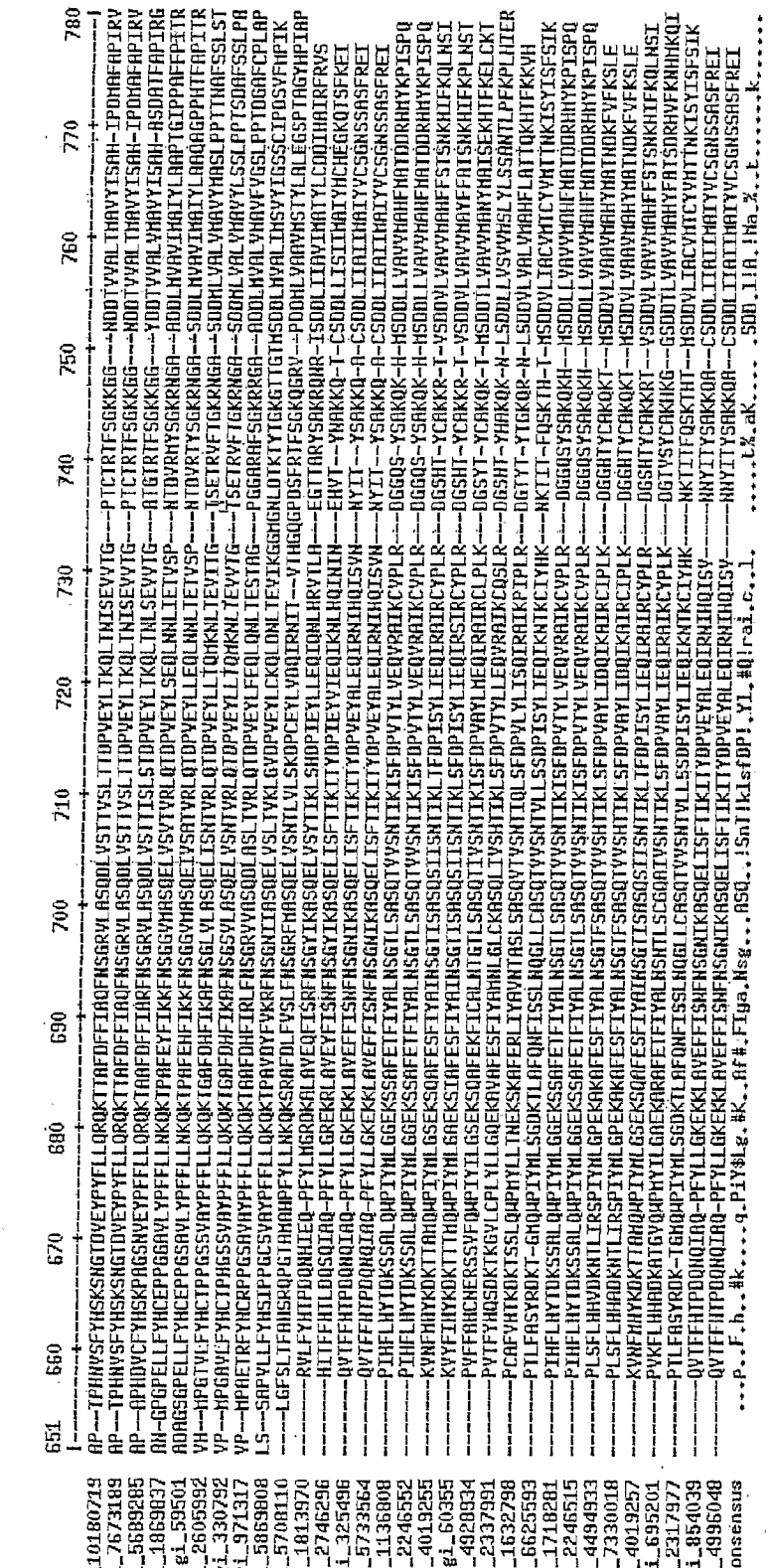


TABLE 2 CONTINUED

```
                    781      791
                    |--------+|
gi_10180719
  gi_7673189
gi_5689285  VEATCRGPTEA
gi_1869837  TS
   gi_59501 VS
gi_2605992  Q
  gi_330792 Q
  gi_971317 RPPAD
gi_5869808
gi_5708110  IGRRQRPA
gi_1813970
gi_2746296
  gi_325496
gi_5733564
gi_1136808
gi_2246552
gi_4019255
   gi_60355
gi_4928934  FF
gi_2337991
gi_1632798
gi_6625593
gi_1718281
gi_2246515
gi_4494933
gi_7330018
gi_4019257
  gi_695201
gi_2317977
  gi_854039
gi_4996048
 Consensus
```

TABLE 2 CONTINUED

Table 3. Degenerate primers generated by CODEHOP

Block x7263xbliD

T L Y V Y I D P
oligo:5'-AACCTGTACGTGtayntngaycc-3' degen=64 temp=33.4 Extend clamp

T L Y V Y I D P A
oligo:5'-AACCTGTACGTGTACntngayccngc-3' degen=128 temp=36.0 Extend clamp

T L Y V Y I D P A Y
oligo:5'-AACCTGTACGTGTACATngayccngcnt-3' degen=128 temp=42.5 Extend clamp

Complement of Block x7263xbliD

Y I D P A Y T N N T
atrnanctrggGCGGATGTGGTTGTTGT oligo:5'-TGTTGTTGGTGTAGGCGggrtcnanrta-3'
degen=64 temp=62.9

D P A Y T N N T R A
anctrggncgnaTGTGGTTGTTGTGGGTCCG oligo:5'-GCCTGGGTGTTGTTGGTGTangcnggrtcna-3'
degen=128 temp=61.8

D P A Y T N N T R A
ctrggncgnawGTGGTTGTTGTGGGTCCG oligo:5'-GCCTGGGTGTTGTTGGTGwangcnggrtc-3'
degen=64 temp=61.0

Block x7263xbliE

C I I F G M E H F F
oligo:5'-TGGATCATCTTCGGCATngarcaytwyt-3' degen=64 temp=55.7 Extend clamp

I F G M E H F F L
oligo:5'-CATCTTCGGCATGGAGcaytwytwyyt-3' degen=64 temp=62.0

Complement of Block x7263xbliE

E H F F L R D L T G
ctygtrawrawGGACTTCCTGGACTGCCC oligo:5'-CCCGTCAGGTCCTTCAGGwarwartgytc-3'
degen=32 temp=61.7

H F F L R D L T G
tygtrawrawrrACTTCCTGGACTGCCCG oligo:5'-GCCCGTCAGGTCCTTCArrwarwartgyt-3'
degen=128 temp=60.8

H F F L R D L T G
gtrawrawrraCTTCCTGGACTGCCCG oligo:5'-GCCCGTCAGGTCCTTCarrwarwartg-3' degen=64
temp=60.8

Block x7263xbliF

E V H I A V E G N
oligo:5'-GGACGTGCACGTCGCCrtngarggnaa-3' degen=64 temp=63.8

Complement of Block x7263xbliF

E G N S S Q D S A
anctyccnttrwGGTTGGTCCTGAGGCGG oligo:5'-GGCGGAGTCCTGGTTGGwrttnccytcna-3'
degen=128 temp=62.7

E G N S S Q D S A V
ctyccnttrwsGTTGGTCCTGAGGCGGC oligo:5'-CGGCGGAGTCCTGGTTGswrttnccytc-3'
degen=64 temp=63.9

CLAIMS

1. A high throughput method for screening a biological sample for unknown viruses, which method comprises

(a) subjecting DNA from the sample to PCR amplification conditions using simultaneously multiple pairs of degenerate primers, wherein each primer binds a sequence that is conserved across members of a family of viruses and each pair of primers selectively directs amplification of sequence of said family;

(b) sequencing PCR product obtained in step (a); and

(c) comparing the sequence of the PCR product with the sequences in at least one database comprising viral sequences to determine whether the sequence is present in, or absent from, the database, wherein absence of the sequence from the database suggests that the sequence may be from an unknown virus.

2. A method according to claim 1 wherein from 12 to 300 pairs of degenerate primers are used simultaneously.

3. A method according to claim 1 wherein from 24 to 96 pairs of degenerate primers are used simultaneously.

4. A method according to claim 1, 2 or 3 wherein the PCR reaction step is carried out in a multi-well plate.

5. A method according to claim 4 wherein the multi-well plate is a 96-well or 384-well plate.

6. A method according to claim 4 or 5 wherein each well contains more than one pair of the degenerate primers.

7. A method according to claim 6 wherein each well contains three pairs of the degenerate primers.

8. A method according to claim 6 or 7 wherein each pair of the primers used in the same well as other pair(s) of primers generates a PCR product of a different size to said other pair(s).

9. A method according to any one of claims 6 to 8 wherein each pair of primers used in the same well as other pair(s) of primers carries a different label from said other pair(s).

10. A method according to claim 9 wherein each pair of primers used in the same well as other pair(s) of primers carries a differently-coloured flourescent label from said other pair(s).

11. A method according to any one of the preceding claims wherein multiple biological samples are screened by simultaneously subjecting DNA from the multiple samples to PCR reaction conditions in step (a) using simultaneously the multiple pairs of degenerate primers on the DNA from each sample.

12. A method according to claim 11 wherein from 2 to 80 samples are screened simultaneously.

13. A method according to claim 12 wherein from 5 to 40 samples are screened simultaneously.

14. A method according to any one of the preceding claims wherein DNA of multiple biological samples is mixed together to produce one or more pooled samples and the PCR reaction of step (a) is carried out on the pooled samples.

15. A method according to any one of the preceding claims wherein the DNA is

genomic DNA.

16. A method according to any one of claims 1 to 14 wherein the DNA is cDNA.

17. A method according to any one of the preceding claims wherein the multiple pairs of degenerate primers are designed by

(i) providing a plurality of amino acid sequences from members of a first virus family,

(ii) comparing the sequences to identify conserved regions,

(iii) designing a first primer pair using a computer based method, wherein each primer in the pair binds a nucleotide sequence that encodes a conserved region identified in (ii) and wherein the primer pair is designed to amplify by PCR the nucleotide sequence between the nucleotide sequences that encode conserved regions in members of the first virus family, and

(iv) repeating steps (i) to (iii) for each virus family.

18. A method according to any one the preceding claims wherein the biological sample is a human sample.

19. A method according to any one of the preceding claims which further comprises determining whether the sequence of the PCR product is a sequence of human DNA.

20. A method according to any one of the preceding claims wherein, if the sequence of the PCR product is absent from the database comprising viral sequences, DNA walking is carried out to determine any sequence which flanks the sequence of the PCR product.

21. A method according to any one of the preceding claims wherein, if the sequence of the PCR product is absent from the database comprising viral sequences, any virus comprising the sequence is isolated.

22. A method according to any one of the preceding claims which further comprises determining whether the sequence of the PCR product or sequence contiguous thereto is present in a specimen from a patient who has a disease, wherein any presence of a said sequence in the specimen suggests that the sequence may be from a virus which is causing or contributing to the disease.

23. A method according to any one of the preceding claims which further comprises

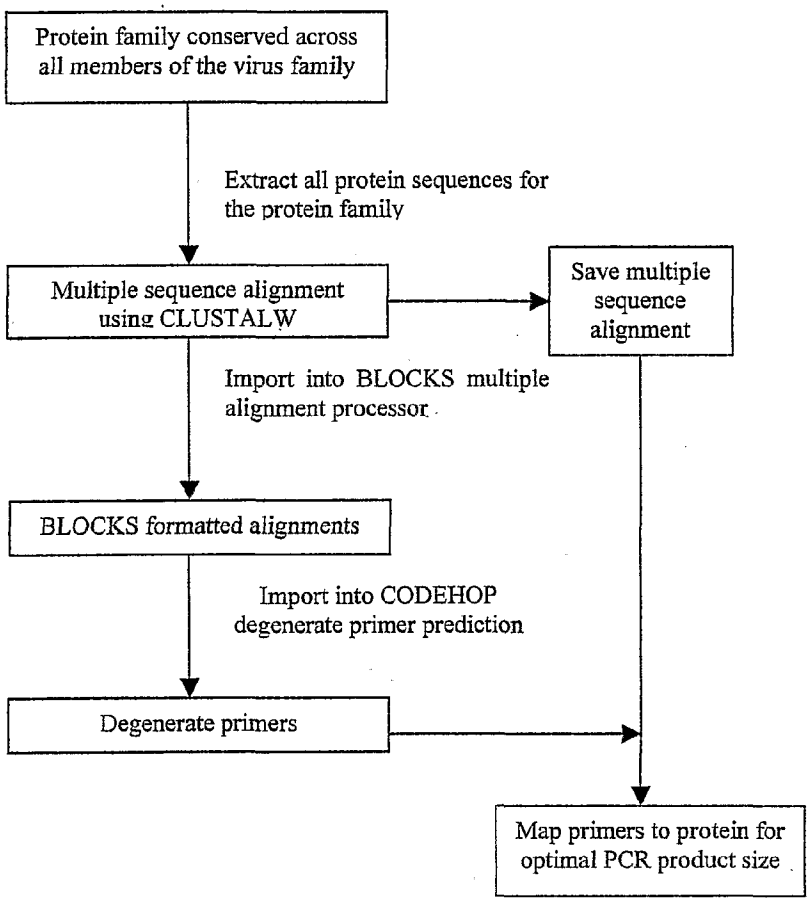
obtaining a specimen from each member of a group of subjects with a disease;

determining whether the sequence of the PCR product or sequence contiguous thereto is present in the specimen from each member of the group; and

determining whether the proportion of subjects in whom a said sequence is present is greater in the group of subjects who have the disease than in a control group of subjects who do not have the disease, wherein a said greater proportion suggests that the sequence may be from a virus which causes or contributes to the disease.

FIG 1

1/1


Protein family conserved across all members of the virus family
Extract all protein sequences for the protein family
Multiple sequence alignment using CLUSTALW
Save multiple sequence alignment
Import into BLOCKS multiple alignment processor
BLOCKS formatted alignments
Import into CODEHOP degenerate primer prediction
Degenerate primers
Map primers to protein for optimal PCR product size